Maastricht University

Gebruik van de Diagnostische Tussentijdse Toets voor opbrengstgericht werken op scholen

Lex Borghans, Ron Diris,
Raoul Haenbeukers en Pascalle Haenen

Universiteit Maastricht
School of Business and Economics
Department of Economics

Inhoudsopgave

1. Inleiding

Op initiatief van het Ministerie van Onderwijs, Cultuur en Wetenschap (OCW) is de afgelopen jaren de Diagnostische Tussentijdse Toets (DTT) ontwikkeld door het College voor Toetsen en Examens (CvTE) in samenwerking met Stichting Cito, DUO en SLO. Het doel van deze landelijk genormeerde toets aan het einde van de onderbouw van het voortgezet onderwijs is om op leerling-, school- en stelselniveau voor de vakken Nederlands, Engels en wiskunde een beeld te geven van hoe de leerlingen er voor staan (OCW, 2011 en Cito, 2012).[1] In september 2014 is begonnen met de ontwikkeling van een driejaarse pilot voor de DTT. Er hebben afnames plaatsgevonden in 2015, 2016 en 2017 waarin de verschillende innovatieve aspecten van de toets in fases zijn uitgerold.[2] Gedurende dit traject is besloten om van overheidswege niet meer verder te gaan met de ontwikkeling van de DTT na afronding van de pilot. Hoewel de DTT in deze vorm niet meer verder wordt afgenomen, zal diagnostisch en formatief toetsen op scholen een belangrijke rol blijven spelen. De inzichten en de gegevens die door de DTT pilot beschikbaar zijn gekomen kunnen een belangrijke bijdrage leveren aan het verder verbeteren van het gebruik van dit type toetsen binnen het onderwijs.

Bij de ontwikkeling van de DTT in de pilot heeft vooral het stellen van diagnoses op leerlingniveau centraal gestaan. Een uniforme toets halverwege het voortgezet onderwijs is echter ook waardevol op een meer geaggregeerd niveau zoals afdelingen van scholen, scholen als geheel, een bepaalde regio in Nederland of Nederland als geheel. Op scholen is er een toenemende aandacht voor een formatieve cultuur waarbij op basis van adequate gegevens gekeken wordt naar de sterke en zwakke punten van de school en waarbij deze gegevens benut worden om de kwaliteit van het onderwijs op peil te houden of te verbeteren. Ook binnen een regio is goede informatie over hoe het onderwijs ervoor staat van belang. Door een goede monitoring kunnen uitdagingen waar het onderwijs in een regio voor staat in beeld worden gebracht, zodat de onderwijsinstellingen gezamenlijk de verantwoordelijkheid hiervoor kunnen nemen. In 2008 liet de commissie Dijsselbloem (Commissie Parlementair Onderzoek Onderwijsvernieuwingen, 2008) zien dat ook voor de verantwoordelijkheid voor het onderwijs op stelselniveau goede informatie over de prestaties van leerlingen van groot belang is.

Op basis hiervan is door de Universiteit Maastricht en de Universiteit Twente onderzoek verricht om na te gaan in hoeverre de DTT,[3] die qua vragen en analyses geoptimaliseerd is om informatie te geven op leerlingniveau, benut kan worden om op geaggregeerd niveau bij te dragen aan het beeld hoe het onderwijs er voor staat. In dit rapport presenteren we onze bevindingen op dit gebied. We benaderen dit onderwerp vanuit verschillende perspectieven. In het eerste deel van het rapport ligt de nadruk op de analytische invalshoek, waarbij we door middel van psychometrische analyses inzichten verschaffen in de implicaties van de verschillende keuzes die gemaakt moeten worden bij het samenstellen en afnemen van toetsen. Vervolgens wordt bekeken hoe de informatie die beschikbaar is op leerlingniveau uit de DTT pilot het beste gebruikt en geaggregeerd kan worden om een informatieve indicator op schoolniveau te construeren. Verder analyseren we wat de implicaties zijn van de gekozen focus voor leerlingniveau in de DTT voor de bruikbaarheid van de indicatoren op

[1] Voor Engels en Nederlands betreft het de schrijfvaardigheid van deze vakken in de pilot afname.
[2] Uitgebreide documentatie over de opzet, implementatie en evaluatie van de pilot is beschikbaar via www.pilotdtt.nl.
[3] Wanneer in dit rapport verwezen wordt naar 'de DTT' dan betreft dit de DTT zoals ontwikkeld in de pilot.

schoolniveau, en in hoeverre een alternatieve opzet de bruikbaarheid op schoolniveau verder zou kunnen vergroten.

Een ander perspectief dat in dit onderzoek naar voren wordt gebracht is dat van de scholen. De betekenis van de DTT als feedbackinstrument op schoolniveau komt uiteindelijk vooral naar voren in de rol die deze toets speelt bij de aansturing van het onderwijs in de praktijk. Om die reden hebben we in verschillende fases van het onderzoek overleg gehad met een aantal scholen over de betekenis die diagnostische en formatief toetsen voor hun schoolbeleid kan hebben. De schoolbezoeken schaffen ook inzicht in hoe de scholen aankijken tegen de keuzes die zijn gemaakt in de opzet van de DTT pilot toets en in de voorwaarden die voor scholen belangrijk zijn om in de toekomst verder te gaan werken met diagnostische instrumenten. Ook zijn de in het kader van dit onderzoek ontwikkelde alternatieven voor het presenteren van informatie uit de DTT voorgelegd aan de gesprekspartners bij de scholen. In deze consultatieronde hebben we met twaalf Limburgse scholen gesproken, waarbij de verschillende onderwijsniveaus waren vertegenwoordigd. De gesprekken zijn gevoerd met de sector-/locatiedirecteur, kwaliteitsmedewerkers en docenten.

Tijdens deze gesprekken over de rol van de DTT in de informatievoorziening van de scholen kwam naar voren dat een landelijk genormeerde en diagnostische toets als de DTT een belangrijke rol kan spelen bij gedifferentieerd en gepersonaliseerd onderwijs. Veel scholen zijn momenteel bezig om dergelijk vormen van onderwijs te ontwikkelen en diagnostische toetsen spelen daarbij een belangrijke rol. Het gaat hierbij niet alleen om het in kaart brengen van leerbehoefte maar ook om de monitoring van voortgang in termen van einddoelen zoals eindexamen. Dit geldt zowel op leerling- als school/afdelingsniveau. Om die reden gaan we in dit rapport ook specifiek in op de betekenis van de DTT voor gedifferentieerd leren.

Het nut op schoolniveau van de informatie uit de DTT kan verder vergroot worden door deze gegevens ook te koppelen aan andere informatie over de betreffende schoolpopulatie. Zo weten we bijvoorbeeld uit eerder onderzoek dat leerlingen met laag opgeleide ouders relatief zwakker scoren op begrijpend lezen. Relatief zwakke schoolscores op dit onderdeel zouden daarmee wellicht voor een deel verklaard kunnen worden door een groot aandeel leerlingen met laag opgeleide ouders. Vanuit een meer algemeen perspectief is het nuttig om te weten met welk niveau de leerlingen instromen op de VO-school, en hoe de prestaties op de DTT zich daartoe verhouden. Op deze manier kan uit elkaar gehaald worden welke verschillen in scores tegenover andere scholen verklaard kunnen worden door achtergrond/instroomverschillen en welk deel is ontstaan door aspecten die direct relateren aan het leerproces op de VO-school. In dit rapport laten we aan de hand van enkele voorbeelden de mogelijkheden zien die ontstaan als toetsinformatie van de DTT wordt gerelateerd aan achtergrondindicatoren en andere toetsresultaten. Dit doen we aan de hand van een koppeling met degelijke gegevens voor een aantal Limburgse scholen. Dit zijn scholen die zowel aan de DTT hebben deelgenomen als periodiek gevolgd worden in het kader van de OnderwijsMonitor Limburg.

Tot slot presenteren we ook een onderwijskundige focus op de DTT, door middel van een literatuurstudie die bekijkt hoe een toets als de DTT van nut kan zijn bij gedifferentieerd leren. Uit deze verkenning van de literatuur komen ook een aantal concrete aanbevelingen naar voren over hoe diagnostische en formatieve toetsen optimaal benut kunnen worden in een tijd waarin leerlingdifferentiatie steeds gangbaarder wordt.

De opbouw van het rapport is als volgt. In Hoofdstuk 2 gaan we in op de principes van het samenstellen en afnemen van toetsen, en de belangrijke rol van meetfout hierin. Hoofdstuk 3 bespreekt de keuzes die gemaakt zijn in de opzet van de DTT, en de consequenties daarvan. In Hoofdstuk 4 introduceren we een alternatieve scoringsindicator op basis van een continue schaal. Hoofdstuk 5 laat vervolgens zien hoe schoolfeedback er uit zou kunnen zien op basis van deze continue score. De meerwaarde van koppeling van prestaties op de DTT aan achtergrondgegevens van leerlingen komt aan bod in Hoofdstuk 6. Hoofdstuk 7 gaat in op de potentie van diagnostische en formatieve toetsen zoals de DTT bij gedifferentieerd leren. De conclusies van dit onderzoek worden gepresenteerd in Hoofdstuk 8.

2. Hoe werkt het samenstellen en afnemen van toetsen?

In dit hoofdstuk wordt kort besproken hoe toetsen de onderliggende vaardigheden van leerlingen meten. We bekijken hoe toetsen op basis van goede en foute antwoorden tot een totaalscore komen, en hoe de nauwkeurigheid van die score afhangt van verschillende factoren, zoals het aantal gestelde vragen en het wel of niet gebruiken van adaptiviteit in de vraagafname. Tot slot bekijken we hoe de verschillende classificaties van toetsen (summatief, formatief, diagnostisch) zich tot elkaar verhouden.

De bespreking van deze onderwerpen wordt ondersteund door een aantal gesimuleerde voorbeelden, die gebruik maken van Item Response Theory (IRT), waarin de vaardigheid die wordt gemeten een continue schaal heeft. Er zijn ook andere gerelateerde methodes die gebruikt kunnen worden om voor een toets een score te bepalen, zoals bijvoorbeeld het latente klassemodel dat in de DTT pilot toets wordt gebruikt. Hierin is de vaardigheid niet continue, maar worden drie niveaus onderscheiden: onder niveau, op niveau en boven niveau. Een continue schaal biedt een goed uitgangspunt om ook alternatieve methodes te bespreken.

2.1 Item Response theory

Het principe van een toets is dat door het voorleggen van een aantal vragen aan (bijvoorbeeld) een leerling een indicatie wordt verkregen van het niveau van zijn of haar kennis op het getoetste onderdeel. Deze indicatie is daarbij dus altijd een inschatting van het werkelijke niveau van de betreffende leerling.

De uitkomst van de toets leidt tot een combinatie van (eventueel deels) goede en foute antwoorden. Er zijn verschillende manieren waarop die goede en foute antwoorden gecombineerd kunnen worden tot een eindoordeel. De meest eenvoudige en een veel gebruikte maat is het aantal goed beantwoorde vragen. Die eenvoudige aanpak heeft echter, ten minste, drie belangrijke nadelen. Ten eerste is het op deze manier moeilijk om de resultaten van twee verschillende toetsen met elkaar te vergelijken. Als twee toetsen niet exact dezelfde vragen hebben kan het aantal goed beantwoorde vragen immers afhangen van zowel de moeilijkheidsgraad van de vragen als het niveau van de leerlingen. Ten tweede is met het aantal goed beantwoorde vragen als maatstaf niet goed mogelijk om vast te stellen hoe groot de nauwkeurigheid van de score is. Om de nauwkeurigheid vast te kunnen stellen is immers een indicatie nodig van de rol die toeval speelt bij de beantwoording van de toets. Ten derde kan via deze methode ook niet bepaald worden welke vragen het beste in een toets kunnen worden opgenomen. Alleen als bekend is welke invloed een additionele vraag heeft op de nauwkeurigheid van de toetsscore kan bekeken worden voor welke leerling welke vraag het beste zou kunnen worden toegevoegd.

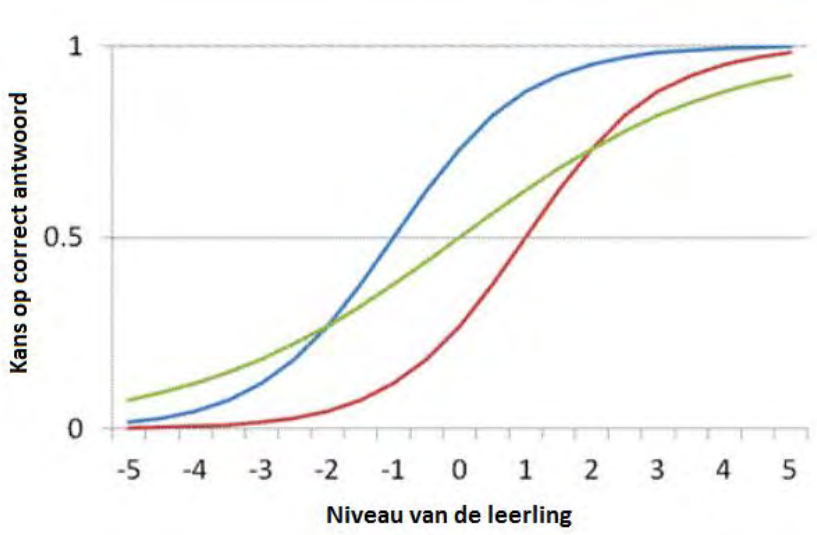
Item Response Theory (IRT) is een aanpak die al deze zaken wel mogelijk maakt. Bij IRT wordt van iedere vraag vastgesteld hoe groot de kans is dat een leerling van een bepaald niveau deze vraag goed beantwoord. Dit is weergegeven in Figuur 1 door de zogenaamde Item Characteristic Curve (ICC). Deze geeft de relatie tussen het niveau van de leerling (op de horizontale as, van laag naar hoog) en de kans dat de specifieke vraag correct wordt beantwoord (op de verticale as).

Voor elke vraag is er één specifieke ICC. De blauwe vraag is makkelijker dan de rode vraag. Leerlingen van eenzelfde niveau hebben immers een grotere kans de vraag goed te beantwoorden.

De blauwe vraag is het meest onderscheidend rond niveau −1. De rode vraag is het meest onderscheidend rond +1. De groene vraag wordt vergeleken met de andere twee vragen vaker goed beantwoord door leerlingen van laag niveau maar minder vaak goed beantwoord door leerlingen van hoog niveau. Deze vraag is dus minder *onderscheidend* in het identificeren van leerlingen van laag niveau en leerlingen van hoog niveau. Hoe steiler de curve, hoe hoger dit onderscheidend vermogen.

De vorm van de ICC-curves in Figuur 1 wordt de 2PL-specificatie genoemd. Het onderscheidend vermogen van de curve is één van de twee centrale parameters van een vraag binnen 2PL IRT. De andere centrale parameter is de moeilijkheid van de vraag. Deze wordt weergegeven door het punt *waar* de helling van de curve het steilst is. Hoe verder naar rechts dit punt ligt, des te moeilijker de vraag. De blauwe en de rode curve hebben dus een vergelijkbaar onderscheidend vermogen maar de rode vraag heeft een hogere moeilijkheidsgraad. De groene vraag zit hier tussen in wat betreft moeilijkheid. In het algemeen geldt dat vragen met een zo hoog mogelijk onderscheidend vermogen de voorkeur hebben. De gewenste moeilijkheid hangt af van veel factoren, maar wanneer de toets afgenomen wordt bij een gevarieerde leerlingpopulatie dan is het van belang dat er een goede mix is van moeilijke en makkelijke vragen. Als bekend is wat het niveau van een leerling ongeveer is, dan zijn vragen die een moeilijkheidsgraad in de buurt van dat niveau hebben optimaal.

Figuur 1: Item Characteristic Curves



Als de ICC-curves voor iedere vraag bekend zijn, dan kan bepaald worden hoe waarschijnlijk het is dat een leerling die bepaalde vragen goed en andere vragen fout beantwoordde een bepaald niveau heeft. Zo wordt duidelijk wat het meest waarschijnlijke niveau is, maar ook hoe breed het interval is van niveaus die ook plausibel zijn. IRT komt tot een uiteindelijke score door verschillende gewichten toe te kennen aan het correct beantwoorden van de verschillende vragen, afhankelijk van moeilijkheid en onderscheidend vermogen. Voor een leerling die al veel vragen goed heeft zal het correct beantwoorden van de blauwe vraag niet sterk meewegen in zijn score, aangezien bijna alle leerlingen van bovengemiddeld niveau deze vraag correct hebben. Het correct beantwoorden van de rode vraag weegt voor deze leerling dan relatief sterker mee. Voor leerlingen van lager niveau werkt dit precies omgekeerd.

Kort samengevat, IRT weegt exact mee *welke* vragen goed en fout worden beantwoord om tot een schatting te komen van het niveau van die leerling.

Zoals eerder aangegeven is een ander voordeel van IRT dat er vergelijkbare schalen gecreëerd kunnen worden tussen verschillende toetsen, waardoor toetsscores door de tijd heen vergeleken kunnen worden, ook wanneer de toets verandert en ontwikkelt. Dit is een cruciaal aspect wanneer we op een betrouwbare manier leerlingen, scholen of stelsels door de tijd heen willen vergelijken.

Om de moeilijkheidsgraad van meerdere vragen op één schaal te krijgen is het niet per se nodig om ze binnen dezelfde toets af te nemen bij een groep leerlingen. Nieuwe vragen kunnen ook worden afgenomen in een toets waarin ook vragen zijn opgenomen die al op deze schaal waren gezet. Hierdoor ontstaat de mogelijkheid om de verzameling vragen voor één schaal te laten groeien zonder dat ze ooit allemaal samen in één toets hebben gezeten. De vragen die al in de eerdere toets zaten, treden hierbij op als een 'anker' voor de continue schaal. Hierdoor ontstaan interessante mogelijkheden. Ten eerste wordt het mogelijk om jaarlijks een toets af te nemen waarbij de vragen deels veranderen, maar er toch een goede vergelijking van het niveau tussen de jaren kan worden gemaakt. Een voorbeeld hiervan is de Centrale Eindtoets basisonderwijs, waarbij de vragen zelf worden vernieuwd, maar ankervragen de vergelijking mogelijk maken. Ten tweede hoeven niet alle leerlingen dezelfde toetsvragen te krijgen om ze toch op één schaal te kunnen beoordelen. Dit wordt bijvoorbeeld gebruikt bij de internationale PISA toets. PISA kent meerdere afnameboekjes met deels verschillende vragen. Door hun onderlinge overlap is een analyse van het niveau van iedere leerling echter toch mogelijk. Doordat er aanzienlijk meer vragen zijn dan een individuele leerling beantwoord, kan er over de prestaties op landenniveau meer gezegd worden dan over de prestaties op leerlingniveau. Ten derde maakt dit principe het mogelijk om adaptief te toetsen. Door leerlingen de vragen te geven die het beste passen bij hun individueel ingeschatte niveau, kan hun individuele prestatie preciezer worden geschat. Door de verdeling van de vragen over leerlingen kan dus gevarieerd worden in de informatieve betekenis van de toets op individueel en op geaggregeerd niveau. In de praktijk wordt een toets vaak geoptimaliseerd voor één niveau. Er kan echter ook een afweging worden gemaakt waarop het informatiebelang op de verschillende niveaus in de samenstelling van de toets wordt meegenomen.

2.2 De relatie tussen meetfout en het aantal gestelde vragen

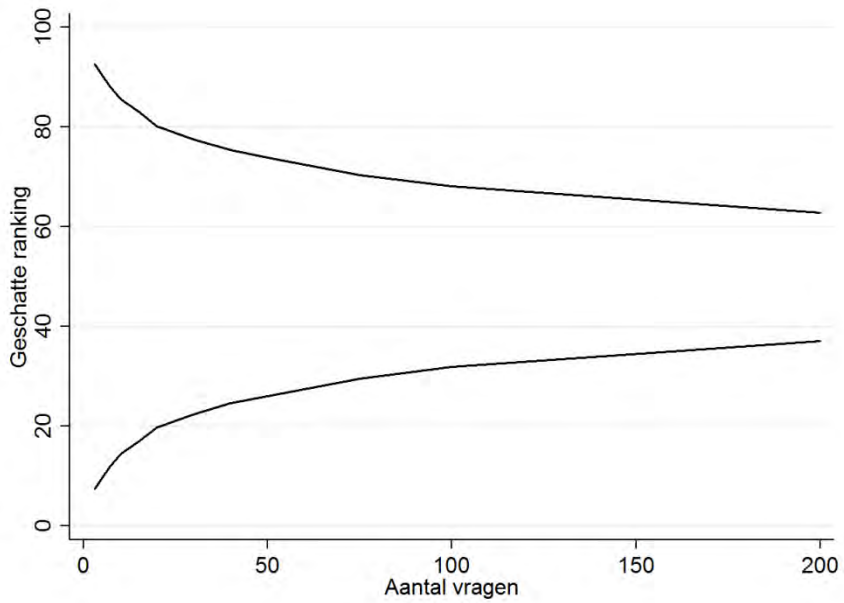
Bij het samenstellen van een toetsscore uit een set aan vragen geldt: hoe meer vragen worden gesteld en hoe geschikter deze voor de betreffende leerling zijn, hoe nauwkeuriger het werkelijke niveau van de leerling bepaald kan worden aan de hand van de toetsscore. De afwijking tussen deze inschatting en het werkelijke niveau van de leerling noemen we de meetfout van de inschatting.

Figuur 2 geeft aan hoe deze relatie tussen nauwkeurigheid en het aantal vragen in elkaar steekt. De parameters die we gebruiken in deze simulatie zijn geschat op basis van de leerlingantwoorden uit de DTT pilot.[4] De figuur laat het 95%-betrouwbaarheidsinterval zien van de inschatting van het niveau van een gemiddelde leerling. Dat betekent dat we met 95% zekerheid kunnen stellen dat het

[4] We schatten allereerst de 2PL IRT parameters van de DTT vragen, uit de adaptieve afname van de 2016 wiskunde toets. Uit het totale spectrum van die parameters trekken we voor deze simulatie willekeurig de parameterwaardes per vraag. De simulatie is gebaseerd op 1800 observaties (ongeveer de grootte van de steekproef in de DTT per onderwijsniveau). Alle simulaties uit dit rapport werken met een 2PL IRT model.

ware niveau van de leerling binnen deze lijnen valt. Scores zijn hier als percentiel uitgedrukt, dus het werkelijke niveau van deze gemiddelde leerling ligt op het 50e percentiel. De bandbreedtes in Figuur 2 centreren zich in alle gevallen ook rond deze 50, maar met een verschillende mate van betrouwbaarheid. We zien dat, wanneer het aantal vragen laag is, de onzekerheid rond dat ware niveau hoog is. Voor iemand die is ingeschat op een gemiddeld niveau bestaat er nog steeds een gerede kans dat hij of zij eigenlijk bij de onderste dan wel de bovenste 20% zit, wanneer het aantal gestelde vragen onder de 20 ligt. De precisie neemt geleidelijk toe met het aantal toetsvragen. Wanneer we informatie hebben over 200 vragen, dan geldt dat we voor iemand die ingeschat is op het 50e percentiel ook met 95% zekerheid kunnen zeggen dat hij tussen het 40e en 60e percentiel zit, en dus werkelijk een 'gemiddeld' niveau heeft. De figuur laat ook duidelijk zien dat de precisie weliswaar toeneemt met meer vragen, maar met afnemende meeropbrengst. De toename in precisie als we van 100 vragen naar 200 vragen gaan is relatief beperkt. Dit is een cruciaal gegeven aangezien het stellen van meer vragen ook kosten met zich mee brengt. Dit is de centrale uitdaging in het samenstellen van toetsen, zoals ook bij de pilot van de DTT: de afweging tussen een grotere informatieve waarde van de uitkomst en meer tijd die nodig is om de toets af te nemen.

Figuur 2: precisie van gemiddelde inschatting naar het aantal gestelde vragen



2.3 De meerwaarde van adaptief toetsen

Onder invloed van technologische ontwikkelingen en de toegenomen digitalisering neemt ook het gebruik van adaptief toetsen in het onderwijs toe. Bij een adaptieve toets worden de vragen die leerlingen krijgen op individueel niveau aangepast aan de prestaties van de leerling tijdens de toets. De achterliggende gedachte hiervan is dat er zo vragen kunnen worden geselecteerd die voor dat type leerling informatiever zijn. Omdat er voor een specifieke leerling 'betere' vragen worden gesteld, wordt de uitkomstmaat preciezer. Voor zwak presterende leerlingen heeft het bijvoorbeeld weinig nut om zeer lastige vragen te stellen als we al weten dat ze dat niveau niet aankunnen. Om te bepalen of een leerling dan een laag niveau of een zeer laag niveau heeft is het veel informatiever om makkelijke vragen te stellen.

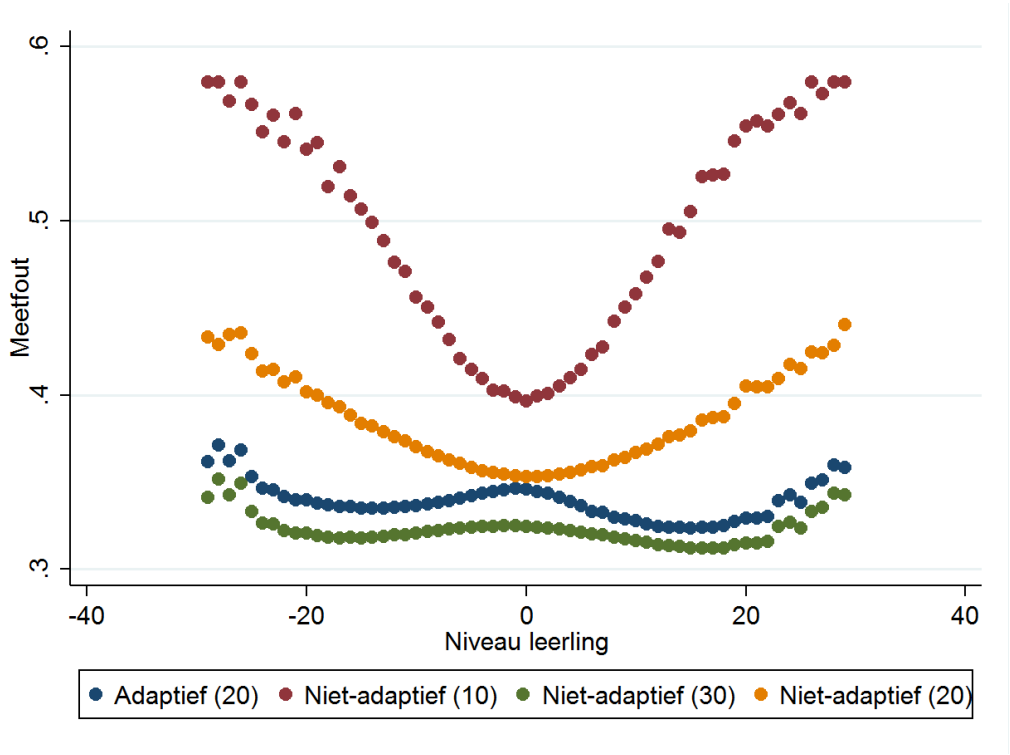
De hogere efficiëntie van een adaptieve toets betekent dus dat bij hetzelfde aantal vragen een hogere precisie bereikt wordt. Anders gezegd betekent het dat voor eenzelfde mate van precisie minder vragen nodig zijn. We bekijken aan de hand van een simpele simulatie wat dit nu concreet uitmaakt voor de precisie van de schatting versus het benodigde aantal vragen, en hoe die winsten behaald worden.

In deze simulatie maken alle leerlingen eerst tien vragen van een gemiddelde moeilijkheidsgraad.[5] Vervolgens krijgen de beter presterende leerlingen er nog tien moeilijke vragen bij en de minder presterende leerlingen krijgen er tien makkelijke vragen bij. We bekijken vervolgens wat de meetfout is van de IRT schatting van het niveau van de leerling. Deze meetfout meet het gemiddelde verschil tussen het gemeten niveau en het werkelijke niveau van de leerling, en is daarmee dus een maatstaf voor de precisie van de schatting. We vergelijken de gemiddelde meetfout in de adaptieve toets met de gemiddelde meetfout van niet-adaptieve toetsen, waarbij we het aantal vragen in de niet-adaptieve toetsen variëren. Op deze manier kunnen we concreet zien hoe de winst door adaptiviteit zich verhoudt tot de winsten die behaald worden door het toevoegen van het aantal vragen en kunnen we zowel zien hoeveel preciezer de adaptieve toets is gegeven hetzelfde aantal vragen, als hoeveel 'sneller' een adaptieve toets is om dezelfde precisie te bereiken.

Figuur 3 geeft dit overzicht. Op de horizontale as zien we het werkelijke niveau van de leerling (van laag naar hoog) en op de verticale as zien we de gemiddelde meetfout. De resultaten voor de adaptieve toets zijn zichtbaar in blauw. Deze zijn vergeleken met de meetfouten voor de toets waarbij we alleen de eerste tien gezamenlijke vragen nemen (in rood) en voor de toets waarbij alle leerlingen alle 30 vragen beantwoorden (de 10 gezamenlijke, de 10 makkelijke en de 10 moeilijke; in groen). Tot slot zien we in het oranje de meetfouten voor een niet-adaptieve toets van 20 vragen (10 gezamenlijke, 5 makkelijke en 5 moeilijke).

[5] Deze simulatie is uitgevoerd op 10,000 (fictieve) leerlingen. De moeilijkheid van de gezamenlijke vragen loopt uniform van -0.9 tot 9, en van de adaptieve vragen ofwel van -2.9 tot -1.1 of van 1.1 naar 2.9. Deze parameters voor de discriminatie van de vragen zijn gebaseerd op schattingen op de afnamegegevens van de adaptieve afname van de DTT in 2016.

Figuur 3: Adaptief toetsen



De adaptieve toets heeft dus eerst dezelfde 10 vragen als alle andere toetsen, en daarna 10 adaptieve vragen. Wanneer we de adaptieve toets vergelijken met de rode toets, dan zien we dat de 10 adaptieve vragen zorgen voor een zeer sterke toename in precisie. Deze toename is met name sterk voor de leerlingen van zeer laag en zeer hoog niveau. Dit is een logisch gevolg van het feit dat de korte niet-adaptieve toets alleen vragen van gemiddeld niveau stelt. De adaptieve toets presteert verder bijna net zo goed als de complete toets van 30 vragen. Het toevoegen van 10 adaptieve vragen na het eerste gezamenlijke blok van 10 vragen is dus bijna even effectief als het toevoegen van 20 niet-adaptieve vragen. De meetfout voor de adaptieve toets is relatief hoger in het midden. Dit komt omdat de adaptieve vragen bovengemiddeld makkelijk (voor de laag presterende leerlingen) of bovengemiddeld moeilijk (voor de hoog presterende leerlingen) zijn, waardoor ze minder onderscheidend zijn voor leerlingen van gemiddeld niveau. Wanneer we de adaptieve toets vergelijken met de niet-adaptieve toets met hetzelfde aantal vragen (in oranje), dan blijkt dat de winst in precisie voor relatief sterke en relatief zwakke leerlingen sterk is, terwijl de toetsen vergelijkbaar presteren voor gemiddelde leerlingen.

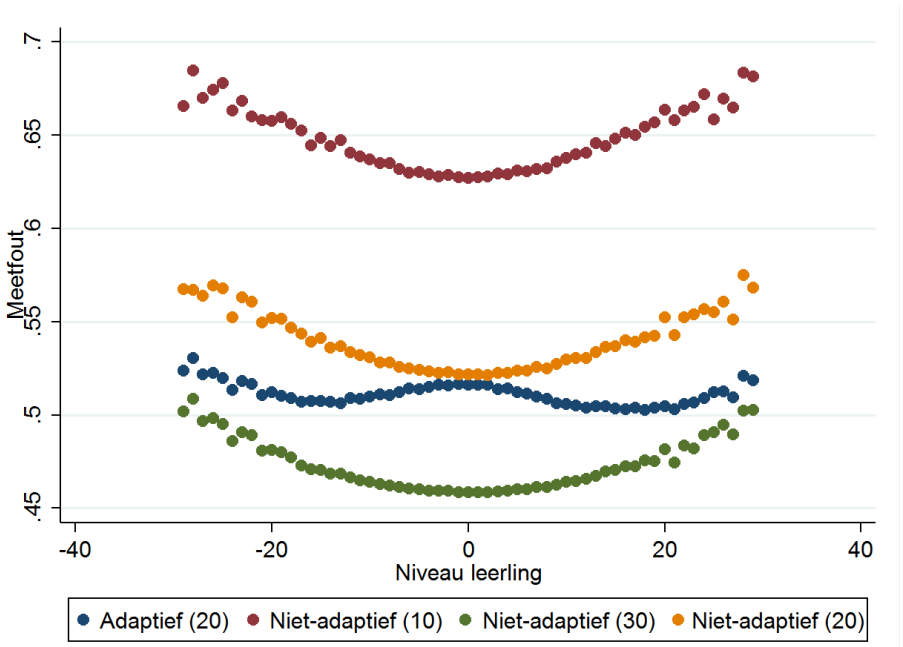
De specifieke patronen in Figuur 3 zijn deels een gevolg van gemaakte keuzes wat betreft de afgenomen vragen. Wanneer we bijvoorbeeld in de niet-adaptieve toets van 20 vragen meer bovengemiddeld moeilijke en bovengemiddeld makkelijke vragen hadden gesteld, dan was het verschil gelijkmatiger geweest. Desalniettemin is duidelijk dat de adaptieve toets gemiddeld beter presteert, en sneller een vergelijkbare mate van precisie haalt. Daarnaast is het een direct gevolg van adaptief toetsen dat er juist minder ‘gemiddelde’ vragen gesteld worden, dus het feit dat de winst sterker is aan de uiteindes van de distributie is een representatief beeld.[6] Aangezien er relatief

[6] Het is hier ook deels een gevolg van de keus voor twee routes. Bij een derde route voor meer gemiddelde leerlingen zou de meetfout daar automatisch wat lager zijn geweest. Tegelijkertijd zou nog steeds gelden dat

meer leerlingen rond het middelpunt zitten dan rond de uiteindes, is het ook een logische keus in niet-adaptieve toetsen om vooral rond dat middelpunt de precisie te maximaliseren. Adaptief toetsen behaalt dus relatief kleine winsten voor die grote middengroep, maar laat wel toe om ook precies te zijn voor die groep daarbuiten, waarvan de onzekerheid rond hun niveau in de traditionele toets juist zo groot is.

Een belangrijke aantekening hierbij is dat de adaptieve toets tegelijkertijd extra afhankelijk is van goede vragen. De winst van adaptief toetsen zit in het feit dat we beter presterende leerlingen vragen kunnen stellen die goed kunnen onderscheiden of een leerling goed of heel goed presteert op dat onderdeel (of zwak tegenover heel zwak). Mocht de onderscheidende waarde van de toets (de hellingshoek in Figuur 1) niet groot zijn, dan valt dat voordeel ook snel weg. In Figuur 4 verzwakken we de onderscheidende waarde van alle vragen in de gesimuleerde toetsen. We zien dat in dat geval het verschil met de toets van 30 vragen groter wordt, en dat de adaptieve toets dichter opschuift naar de niet-adaptieve toets van 20 vragen (maar wel nog steeds preciezer is). De kwaliteit van de vragen is voor een adaptieve toets dus extra belangrijk.

Figuur 4: Adaptief toetsen met slechtere vragen



Bij de DTT pilot is niet gekozen voor een continue score, maar worden de resultaten via een Latent Klassemodel gepresenteerd in de vorm van drie categorieën: “onder niveau”, “op niveau” en “boven niveau”. De psychometrische aanpak is echter vergelijkbaar. In plaats van een ICC per vraag, wordt nu per vraag voor deze drie klassen vastgesteld hoe groot de kans is op een goed antwoord. Afhankelijk van welke vragen een leerling goed of fout beantwoordt kan de kans worden bepaald dat een leerling tot de eerste, tweede of derde klasse behoort. In de DTT worden voor de diagnoses van de hoofdaspecten drie blokken met vragen gesteld, tenzij de diagnose al eerder zeker genoeg is. Voor de diagnoses van de deelaspecten wordt er per deelaspect een extra blok aan vragen

de winst van de adaptieve toets in het midden relatief kleiner is, omdat niet-adaptieve toetsen meer op deze leerlingen zijn afgestemd.

voorgelegd bij een onzekere diagnose, mits er nog voldoende toetstijd is. In feite is het doel van de adaptiviteit dus niet primair om nauwkeurigheid te maximaliseren of om toetstijd te minimaliseren, maar om zo veel mogelijk (zekere) diagnoses te stellen. Toch gelden hierbij dezelfde principes als in de besproken voorbeelden. Betere vragen zullen leiden tot meer diagnoses en het eerder bereiken van de gewenste niveaus van zekerheid voor die diagnoses, wat vervolgens weer tijd vrijmaakt om op andere deelaspecten extra vragen te stellen. Bovendien speelt ook hier wederom het spanningsveld tussen de informatieve waarde van de uitkomst (in dit geval het aantal gestelde diagnoses) en de benodigde toetstijd hiervoor.

2.4 Summatief versus formatief toetsen

Toetsen kunnen gebruikt worden voor meerdere doeleinden. Binnen het Nederlandse onderwijssysteem hebben toetsen traditioneel een *summatieve* functie gehad, waarbij het doel is om niveau en kennis van leerlingen te meten aan het einde van een bepaald traject. De Centrale Eindtoets PO en het Centrale Eindexamen zijn twee typische voorbeelden hiervan. De laatste jaren is er steeds meer aandacht voor de rol van *formatief* toetsen binnen het onderwijs. Waar een summatieve toets vooral gericht is op het verleden, is het doel van een formatieve toets primair om te kijken waar leerprocessen bijgestuurd kunnen worden richting de toekomst. Formatieve toetsen kunnen daarbij zowel voor leerlingen als voor leraren en scholen aandachtspunten bloot leggen.

In de praktijk is het onderscheid tussen formatieve en summatieve toetsen niet altijd even duidelijk als in de theorie wordt uiteengezet. Leren is een continue proces en wat geleerd wordt in de ene periode heeft een sterke relatie met wat geleerd wordt in eerdere en in latere periodes. In de praktijk zijn er daarmee in feite geen afgebakende leerperiodes en kan elke 'summatief' bedoelde toets in principe ook gebruikt worden als een formatief instrument voor zowel leerlingen als scholen om leerprocessen in de toekomst bij te sturen. Tegelijkertijd kunnen formatieve toetsen ook altijd gebruikt worden om summatieve conclusies uit te trekken. Of een toets een formatieve of een summatieve functie vervult, hangt daarom vooral af van hoe een school, leraar of leerling deze gebruikt.

Desalniettemin maakt het voor de ontwikkeling van een toets wel uit of deze (vooral) bedoeld is voor formatieve doeleinden of voor summatieve doeleinden. De DTT is ontworpen voor formatieve doeleinden. Dit vertaalt zich in de diagnostische opzet van de toets, waarbij op een zeer gedetailleerd niveau binnen bepaalde vakken wordt gekeken op welke deelaspecten van die vakken een leerling beter of slechter presteert, terwijl bij een summatieve toets meestal alleen wordt gekeken naar een totaalscore. Een goede formatieve toets vereist een dergelijke opzet in deelaspecten, zodat gericht gewerkt kan worden aan specifieke tekortkomingen.

2.5 Diagnostisch toetsen

Het formatieve perspectief van de DTT vertaalt zich dus in het diagnostische aspect van de toets. Door middel van de diagnoses van de toets kunnen de resultaten gebruikt worden om het leerproces aan te passen. Diagnostisch toetsen brengt in de praktijk vaak met zich mee dat er binnen vakken specifiek op bepaalde onderdelen wordt ingezoomd.[7] Dit is geen absolute vereiste van

[7] Voor een overzicht van de verschillende typen van diagnostische toetsen die onderscheiden worden, zie Rupp et al. (2010). Kenmerkend aan de DTT binnen de verschillende types van diagnostische toetsen is dat het gebaseerd is op een cognitief model, of leerlingmodel.

diagnostisch toetsen. Een school kan immers ook alleen uitkomsten voor wiskunde als geheel gebruiken om vervolgens, bijvoorbeeld, leerlingen op te delen in groepjes op basis van die algehele prestatie. Als in de praktijk echter blijkt dat problemen met wiskunde vaak geconcentreerd zijn bij bepaalde onderdelen van wiskunde, dan is dit een vrij ruwe aanpak van dat probleem. Om effectief te kunnen werken aan juist die onderdelen van wiskunde die extra aandacht verdienen, is een diagnose op een meer specifiek niveau dus nuttig. Hóe nuttig dat is, hangt af van hoe sterk deelvaardigheden voor een bepaalde leerling kunnen verschillen. Komen leerlingen met lagere algehele wiskundeprestaties tekort op alle deelaspecten van wiskunde, of zien we juist vaak combinaties van relatief slechte en relatief goede deelaspecten voor een bepaalde leerling? Sectie 3.5 bekijkt dit specifiek. In ieder geval geldt dat een diagnose op een meer specifiek niveau betere mogelijkheden biedt om ook specifieke onderliggende leerproblemen aan te pakken.

Wanneer er inderdaad gekozen wordt voor een aanpak waarop er binnen een bepaald schoolvak wordt ingezoomd op onderdelen, blijft de vraag hoe dit wordt vormgegeven. Een eerste vraag ligt in de ‘diepte’ van de keuze; op hoeveel niveaus willen we een verdere uitsplitsing maken? Dit zal deels een psychometrische vraag zijn en deels een conceptuele. Psychometrisch gezien geldt dat hoe meer we inzoomen, hoe kleiner de set aan vragen die dat onderdeel meet en hoe lager de precisie (gegeven een bepaalde toetstijd). Belangrijk is dat er op het laagste niveau van onderscheid nog steeds een voldoende mate van precisie is om met voldoende zekerheid uitspraken te doen over het niveau op dit onderdeel. We bekijken later in dit rapport hoe het zit met de nauwkeurigheid op dit laagste niveau binnen de DTT (op leerlingniveau en op schoolniveau), en hoe deze afhangt van bijvoorbeeld keuzes rondom het aantal gekozen onderdelen. Hier speelt dus ook weer een mogelijke afweging tussen het individuele perspectief en het schoolperspectief. Omdat we op schoolniveau kunnen aggregeren en daardoor preciezere uitkomsten krijgen, kan de uitsplitsing vanuit dat oogpunt dus verder doorgevoerd worden. Tegelijkertijd kan de precisie op individueel niveau bij een dergelijke diepe uitsplitsing te laag zijn om nog betekenisvolle conclusies te trekken.

Conceptueel gezien zijn er ook afwegingen. Het inzoomen op onderdelen kan in principe zeer ver gaan. Binnen wiskunde kan specifiek gekeken worden naar meetkunde, waarbinnen weer specifiek gekeken kan worden naar driehoeken, waarbinnen weer specifiek gekeken kan worden naar het berekenen van hoeken van driehoeken etc. De vraag is alleen of het, wanneer de toetsresultaten gebruikt worden om leerprocessen in de klas bij te sturen, effectief is om leerlingen specifiek aan opdrachten te laten werken waarin ze hoeken binnen driehoeken moeten berekenen, en daarna een andere even specifieke opdracht etc. Wellicht is het effectiever om extra aandacht te besteden aan de algehele principes van meetkunde. In dat geval is te veel specificiteit weinig informatief. Zulke aspecten zullen bekeken moeten worden in samenwerking met zowel onderwijskundigen als scholen zelf. In Sectie 3.6 gaan we in op de feedback die we op dit gebied hebben gekregen van scholen.

Een vergelijkbare keuze ligt in de ‘breedte’ van de specificatie. Hoe veel onderdelen van wiskunde willen we in eerste instantie onderscheiden, en hoe veel onderdelen daarbinnen? Hier spelen dezelfde afwegingen als in de keuze voor de mate van detail. Hoe breder de uitsplitsing, hoe kleiner de set aan bijbehorende vragen (of hoe meer de benodigde toetstijd). Daarnaast spelen ook hier weer dezelfde conceptuele afwegingen, in hoe nuttig het voor scholen is om een breed uitgesplitste set aan domeinen te hebben.

Een belangrijke vraag is verder voor welke *combinatie* van breedte en diepte gekozen wordt. Wordt er bijvoorbeeld gekozen voor een opsplitsing van vier hoofdaspecten met elk weer drie deelaspecten of voor een opsplitsing van zes hoofdaspecten met elk twee deelaspecten.[8] In beide gevallen zijn er op het laagste niveau twaalf deelaspecten en is er in principe een gelijke toetstijd nodig om op dit niveau nog voldoende mate van precisie te krijgen. Anders gezegd zal de mate van precisie op het deelaspect bij een gelijke toetstijd vrijwel gelijk liggen in beide scenario's. Een voordeel van de eerste aanpak is echter dat de hoofdaspecten preciezer geschat zullen worden, aangezien ze gegevens combineren van drie deelaspecten, tegenover twee deelaspecten in het andere scenario. Aan de andere kant is in scenario 2 weliswaar de precisie op hoofdaspect lager, maar zijn er wel *meer* hoofdaspecten waarover we informatie hebben. Dit kan vooral een voordeel zijn wanneer het (bijvoorbeeld door een beperkte toetstijd) moeilijk haalbaar is om op deelaspect veel precisie te krijgen. In dat geval zijn er meer onderdelen waarover we wel voldoende zekerheid hebben in scenario 2 dan in scenario 1. De keuze hangt ook weer af van de behoefte van scholen. Hebben zij vooral meer behoefte aan meer onderscheid op het eerste niveau van specificatie, of eerder aan meer onderscheid binnen die hoofdaspecten. Zie ook hiervoor de discussie in Sectie 3.6.

Andere keuzes die gemaakt worden in de opzet van een toets liggen in hoe de uitkomst wordt vormgegeven. Dit geldt onder meer voor de keuze voor een continue of een categoriale indicator en voor de keueze voor een relatieve of een absolute grenswaarde voor de categorieën. Dit zijn echter geen aspecten die bepalend zijn voor de diagnostische functie van een toets. Deze keuzes zullen daarom besproken worden in andere delen van dit rapport.

Tot slot is het belangrijk om te benadrukken dat een diagnostische toets niet op zichzelf kan staan. Diagnostische toetsen zijn ontwikkeld om het leerproces bij te sturen. Het is daarbij belangrijk dat deze aanpassingen ook vervolgens geëvalueerd kunnen worden, zowel op leerlingniveau als op schoolniveau. Een diagnostische toets wordt dus idealiter meermaals afgenomen en maakt daarbij onderdeel uit van een continue formatief proces van evaluatie en optimalisatie. Hierbij moeten ook keuzes gemaakt worden wat betreft de tijd tussen de opeenvolgende toetsmomenten en de onderdelen die daarbij getoetst worden. We bespreken dit verder in de Hoofdstukken 7 en 8.

Conclusie

In dit hoofdstuk is kort beschreven hoe met een IRT model toetsen worden geanalyseerd en hoe dit afhangt van bepaalde parameters en keuzes. Een cruciaal aspect van elke toets is dat de onderliggende vaardigheid altijd met een bepaalde onnauwkeurigheid wordt gemeten. Deze onnauwkeurigheid kan verkleind worden door meer vragen te stellen (waarbij de meeropbrengst gestaag afneemt), betere vragen te stellen en door adaptiviteit in de vraagafname toe te passen. In het laatste geval wordt de vraag die een leerling krijgt, aangepast aan de inschatting van zijn niveau tot dan toe. Er zijn verschillende methodes om een uiteindelijke score te meten, waarbij we hier het IRT 2PL model als leidraad nemen, maar de beschreven principes werken op een vergelijkbare manier voor andere psychometrische aanpakken.

Door het ontwerp van een toets kunnen de meetfout en lengte van de toets worden beïnvloedt. Afhankelijk van het doel van de toets zal deze afweging verschillen. Een optimaal ontwerp voor een

[8] We refereren in de rest van dit rapport naar de eerste laag van uitsplitsing als hoofdaspecten en naar de tweede laag van uitsplitsing als deelaspecten.

toets die scores op leerlingniveau moet bepalen is anders dan een optimaal ontwerp voor een toets voor scores op schoolniveau. Om een toets meerdere functies te geven kan bij het ontwerp echter ook een afweging gemaakt worden tussen de verschillende doelen.

Een ander onderscheid dat traditioneel wordt gemaakt is tussen summatieve toetsen, die bedoeld zijn om een niveau te meten aan het einde van een leertraject, en formatieve toetsen, die gericht zijn op bijsturing van het toekomstige leertraject. In de praktijk gaat het onderscheid tussen formatief en summatief vaak over de vormgeving van de einduitkomst. Dit is echter niet wat een toets summatief of formatief maakt. Het gaat erom hoe de uitkomst wordt ingezet richting het leerproces. Daarbij is het bij een formatieve toets wel cruciaal dat er op een gedetailleerd niveau gekeken wordt waar verbeterpunten (of juist sterke punten) liggen, en niet alleen naar een algehele gemiddelde score.

3. De opzet van de DTT

In de DTT pilot zijn verschillende keuzes gemaakt over de opzet van de toets. Er moest besloten worden hoeveel onderdelen er getoetst worden en welke en wat voor items hierbij afgenomen moeten worden. Ook zijn er keuzes gemaakt in hoe de adaptiviteit in de afname van de vragen wordt ingepast. In dit hoofdstuk beschrijven we deze keuzes en hun implicaties, en hoe deze zich verhouden tot alternatieve keuzes die gemaakt hadden kunnen worden.

Verder analyseren we aan de hand van de afnamegegevens hoe de 'diagnoses' uit de DTT zijn verdeeld. Een centraal aspect van de DTT is dat er binnen vakken op deelaspecten ingezoomd kan worden. De waarde hiervan zal vooral groot zijn als het vaak voorkomt dat achterstanden binnen een vak of binnen een hoofdaspect van dat vak in een deelaspect zijn geconcentreerd. We bekijken in hoeverre dergelijke verschillen tussen deelaspecten binnen een onderdeel in de praktijk blijken voor te komen, en wat dit voor implicaties heeft voor de opzet en evaluatie van diagnostische toetsen.

3.1 Onderdelen van de toets

De DTT is afgenomen voor de vakken wiskunde, Nederlands, en Engels. De toets meet de algemene prestatie op deze vakken, maar ook de prestatie op deelaspecten van elk vak. Elk vak is, in de meest uitgebreide versie, onderverdeeld in ofwel een viertal ofwel een vijftal hoofdaspecten, die elk weer verder onderverdeeld worden in twee of drie (en in een enkel geval vier) deelaspecten.[9]

In Tabel 1 en 2 worden al deze hoofdaspecten en deelaspecten uiteengezet. Voor de wiskundetoets is er een verandering als we de 2016 toets vergelijken met de 2017 toets. In beide jaren zitten de onderdelen 'Meten en Meetkunde' en 'Verbanden en Formules', maar de 2017 toets bevat daarbij nog drie andere hoofdaspecten (deze zijn in de 2016 pre-test eerst uitgetest). Een ander typisch aspect van de wiskundetoets is dat de deelaspecten voor elk hoofdaspect dezelfde aanduidingen heeft: structuur, meerduidigheid en samenhang. Voor de specifieke definities van deze deelaspecten verwijzen we naar de toetswijzer van het CvTE (College voor Toetsen en Examens, 2014).

[9] Het domein 2.4 is alleen afgenomen voor havo- en vwo-leerlingen

Tabel 1: Onderdelen van de vakken Nederlands en Engels

Nederlands	**1-Afstemmen op doel en publiek**	**2-Tekststructuur**	**3-Woord- en zinsniveau**	**4- Spelling en Interpunctie**
	1.1- Voorkennis en informatievoorziening inschatten bij lezer	2.1-Tekstelementen kiezen, rekening houdend met het genre	3.1-Correcte zinsbouw hanteren	4.1-Werkwoorden correct spellen
	1.2- Toonzetting op lezer afstemmen	2.2-Juiste volgorde, indeling en lay-out in teksten aanbrengen	3.2-Passende schrijfstijl hanteren en samenhang op zinsniveau aanbrengen	4.2-Overige regelgeleide spelling correct toepassen
	1.3- Schrijfdoel bepalen	2.3-Samenhang tussen tekstelementen aanbrengen	3.3-Passend en gevarieerd woordgebruik laten zien	4.3-Leestekens en hoofdletters correct hanteren
		2.4-Standpunt weergeven en van passende argumenten voorzien		
Engels	1-Afstemmen op doel en publiek	2-Samenhang	3-Woordenschat en woordgebruik	4-Grammatica, spelling en interpunctie
	1.1-Toonzetting en register afstemmen	2.1-Tekststructuur en verbanden aanbrengen	3.1-Passende woorden en woordcombinaties gebruiken	4.1-Woordvolgorde en zinsconstructie functioneel hanteren
	1.2-Conventies bij tekstsoort gebruiken	2.2-Passende structuurwoorden gebruiken: voegwoorden en verwijswoorden	3.2-Woordgebruik functioneel variëren	4.2-Passende spelling en interpunctie hanteren

Tabel 2: Onderdelen van het vak wiskunde (dikgedrukt komt voor in alle jaren)

B-Getallen	**C-Verhoudingen**	D-Meten en meetkunde	E-Verbanden en formules	**F-Informatieverwerking en onzekerheid**
B1-Structuur	C1-Structuur	**D1-Structuur**	**E1-Structuur**	F1-Structuur
B2-Meerduidigheid	C2-Meerduidigheid	**D2-Meerduidigheid**	**E2-Meerduidigheid**	F2-Meerduidigheid
B3- Samenhang	C3- Samenhang	**D3- Samenhang**	**E3- Samenhang**	F3- Samenhang

3.2 De type items van de DTT

De DTT is een technisch zeer geavanceerde toets. Een vaak voorkomend nadeel van gestandaardiseerde of 'centrale' toetsen is dat ze beperkend zijn in het type vragen dat gesteld kan worden, met name vanuit praktische overwegingen. Dit laat bijvoorbeeld niet makkelijk toe om open vragen met veel tekst als antwoord of wiskundige vragen met ingewikkelde afleidingen af te

nemen. Voor digitale toetsen is die beperking vaak nog sterker, aangezien bijvoorbeeld het evalueren van essays software-technisch nog steeds erg lastig is, terwijl veel wiskundevragen over geometrie of formules beter werken met pen en papier dan digitaal. Een probleem bij digitale wiskundevragen is ook dat er vaak een veelvoud aan antwoorden correct kan zijn, die dan allemaal door de software herkend moeten worden. Gestandaardiseerde en digitale toetsen leunen daarom vaak sterk op multiple choice vragen of korte invulvragen. De DTT gaat verder dan die traditionele aanpak, en gebruikt een combinatie van die meer traditionele vraagtypes en meer innovatieve vraagtypes.[10]

De wiskundetoets bestaat voor het merendeel uit open vragen. Het innovatieve van de toets zit vooral in de vragen over geometrie. Leerlingen moeten daarbij bijvoorbeeld punten aangeven binnen het raster om lijnen of figuren te vormen. Daarnaast bestaan veel vragen uit het invullen of afleiden van formules.

De Nederlands en Engels toetsen gebruiken een grote diversiteit aan type vragen: multiple choice, multiple response, slepen van tekst, ordering van paragrafen, aanwijzen van fouten in de tekst (eventueel met opgave voor verbetering), en invuloefeningen. Het was in de analyses voor dit rapport niet mogelijk om te analyseren welke type vragen het meest bijdragen aan de informatieve waarde van de toets, omdat de informatie over het type vraag per item hiervoor ontbreekt. Voor toekomstig onderzoek kan het interessant zijn om op basis van deze informatie een meer systematische analyse te doen van welke type vragen relatief beter zijn.

Uit een algehele analyse van de kwaliteit van de vragen blijkt dat deze relatief het hoogst is voor Engels, en relatief het laagst voor wiskunde. Verder blijkt de moeilijkheidsgraad van de vragen vrij gelijkwaardig verdeeld. Vanuit het perspectief van het Latente Klassemodel zou het interessant zijn geweest om vooral vragen te selecteren die goed onderscheidend zijn rond de grenswaardes. Dit is in de praktijk niet zichtbaar, waarschijnlijk omdat er in de overgang van pre-test naar adaptieve toets vooral is gekozen om vragen met een algehele slechte kwaliteit uit te sluiten. Voor toekomstige ontwikkelingen van diagnostische toetsen gebaseerd op een Latente Klassemodel kan het een interessante overweging zijn om vooral deze vragen meer toe te voegen aan de gehele itembank.[11]

3.3 Welke keuzes zijn gemaakt in het inpassen van adaptief toetsen in de DTT?

In de DTT wordt er in de adaptiviteit ook gewerkt met een blokschema, enigszins vergelijkbaar met het beschreven voorbeeld in Sectie 2.3. Alle leerlingen krijgen dezelfde beginvragen waarna er, afhankelijk van de antwoorden, tot twee keer toe overgegaan wordt naar één van drie nieuwe blokken. Wanneer er na de drie blokken van vragen niet genoeg zekerheid is over het niveau van de leerling (en er voldoende tijd is in de toets), worden er nog een aantal extra vragen voor dat deeldomein gesteld, totdat de gewenste zekerheid bereikt is (of totdat de vragen voor dat blok of de tijd op zijn).[12] Deze adaptieve opzet is pas vanaf de derde afname (in 2017) op deze manier gedaan;

[10] Zie voor meer informatie College voor Toetsen en Examens (2014).
[11] Als ten minste wordt gekozen voor een focus op de leerling, aangezien dit voor de analyse op schoolniveau weer nadelen kan hebben; zie Sectie 5.2.
[12] Voor een meer uitgebreide uitleg van de adaptiviteit in de DTT verwijzen we door naar de uitgebreide documentatie op de DTT Pilot website.

in de adaptieve afname van 2016 is eerst nog besloten om een wat beperktere mate van adaptiviteit in te bouwen, als opbouw naar het complete model.

In het gesimuleerde voorbeeld in Sectie 3.2 werd de route van de adaptieve toets simpelweg bepaald door te kijken naar het aantal goede antwoorden. Recente toetsen die gebruik maken van adaptiviteit gebruiken daarvoor meer geavanceerde algoritmes. Een vaak gekozen aanpak in adaptief toetsen is om het algoritme te kiezen dat ofwel de nauwkeurigheid maximaliseert ofwel de toetstijd minimaliseert. De DTT aanpak is engiszins verschillend. Het doel van de DTT is om zo veel mogelijk diagnoses te stellen. Diagnoses worden gesteld wanneer er in het latente klasse model een niveau van zekerheid wordt bereikt dat impliceert dat minstens 90% van de leerlingen een correcte diagnose zou krijgen. De adaptiviteit is ingezet zodat de vragen geselecteerd worden die het vaakst zorgen dat die grenswaarde bereikt wordt.

De aanpak van de DTT is dus meer geavanceerd dan het simpele voorbeeld in de vorige sectie. Hoe veel levert deze toegepaste vorm van adaptiviteit nu concreet op in de DTT als het gaat om hogere precisie en mindere toetstijd? Om dit te concretiseren lijkt het in eerste instantie logisch om te kijken naar de verschillende versies van de DTT, waarin de adaptiviteit langzaam is uitgerold.[13] Deze vergelijking is in de praktijk echter moeilijk, omdat ook andere aspecten verschillen tussen de versies. Zo kennen de pre-toetsen ten opzichte van de adaptieve toetsen een langere toetstijd aangezien er ook meer vragen in zijn beantwoord. Daarnaast is het ook lastig om de adaptieve afname van 2016 te vergelijken met de adaptieve afname van 2017 (waarin de adaptiviteit sterk is doorontwikkeld) aangezien de 2017 versie een veel bredere verzameling aan mogelijke vragen kent én een hogere effectieve toetstijd heeft. De mate van precisie is uiteindelijk hoger bij de 2017 afname, maar dit is met name zichtbaar bij domeinen waar ook veel meer vragen voor zijn afgenomen. Voor domeinen waarvoor de toetstijd weinig is veranderd is de toename in precisie relatief kleiner.

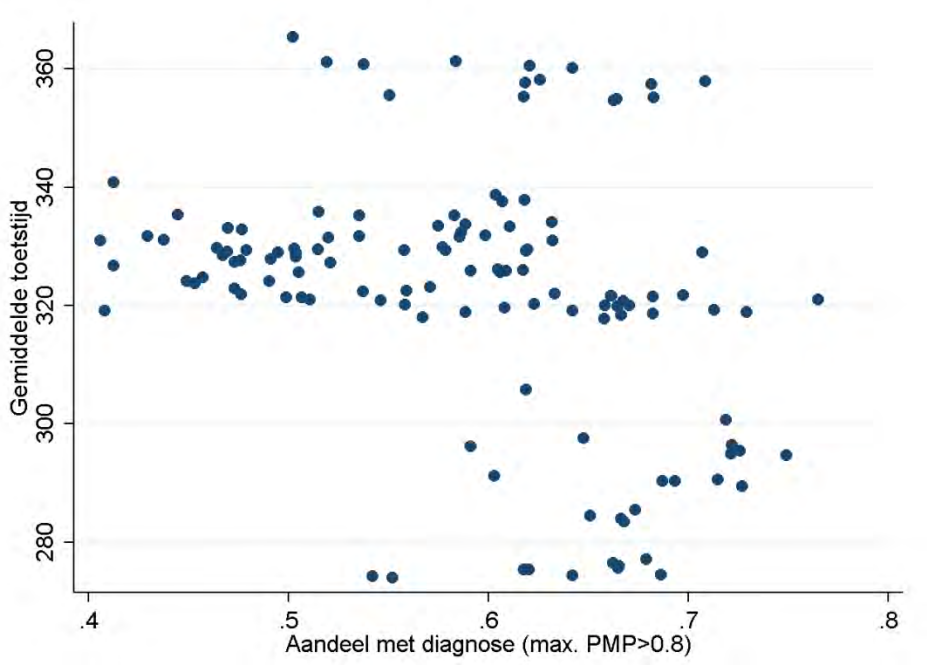
In ieder geval levert de adaptiviteit een bepaalde winst op in precisie. Buiten het kiezen voor wel of geen adaptiviteit kan er ook nog een keuze worden gemaakt in hoe de adaptiviteit wordt ingepast. De centrale keuze in de huidige opzet is om te kijken welke vragen de kans op een diagnose maximaliseren. Een andere optie is om het algoritme (ook) af te laten hangen van de toetstijd. We analyseren een simpel voorbeeld waarbij we uit een totale verzameling van 9 vragen steeds 5 vragen selecteren. Ter vereenvoudiging zijn er in het voorbeeld slechts twee niveaus; onder niveau of boven niveau. We specificeren verder in dit gesimuleerde voorbeeld dat de toets afgekapt wordt wanneer we na een vraag voor ten minste 80% zeker weten dat de leerling in één van beide categorieën valt. Hoe sneller we dit punt bereiken, hoe korter de toetstijd. Wanneer dit punt niet bereikt wordt, stopt de toets na vijf vragen. Er zijn een veelvoud aan combinaties en volgordes van de verschillende vragen, die allemaal een bepaalde combinatie hebben van gemiddelde toetstijd en een aandeel gediagnostiseerde leerlingen (aandeel leerlingen waarbij de Posterior Model Probability (PMP) boven de 80% uitkomt).[14]

[13] De eerste afname van de DTT pilot, in 2015, was een zogenaamde 'pre-toets', waarin standaarden werden gezet voor de adaptieve afnames in de jaren erna. In 2016 vond zowel een adaptieve afname als een nieuwe pre-toets (om nieuwe vragen te toetsen voor de afname van 2017) plaats.

[14] Deze simulatie is gebaseerd op 2000 observaties. De IRT parameters zijn net als voor Figuur 2 willekeurig getrokken uit het spectrum van geschatte parameters van de DTT vragen.

In Figuur 5 zien we de beide parameters voor al deze combinaties. Uit ons gesimuleerd voorbeeld blijkt dat de combinatie van vragen die het aantal zekere diagnoses maximaliseert (het meest rechtse datapunt) niet dezelfde combinatie van vragen is die ook de toetstijd minimaliseert (het laagste datapunt). Laatstgenoemde combinatie laat dus toe om voor leerlingen met een lage zekerheid nog extra vragen te stellen. Wanneer we voor de combinaties met de lage toetstijd nog een extra (zesde) vraag stellen, levert dit voor sommige combinaties een totaal aantal diagnoses op dat hoger ligt dan het maximum in Figuur 5, terwijl de gemiddelde toetstijd nog steeds onder dat punt blijft liggen. Het meenemen van tijd in het algoritme kan dus verdere winsten betekenen in zowel nauwkeurigheid diagnoses als toetstijd (en dus ook in het aantal gestelde diagnoses).

Figuur 5: zekerheid en toetstijd per exacte combinatie van vragen



3.4 Welke keuzes zijn gemaakt in het inpassen van het diagnostische aspect?

In de keuze van het vormgeven van het diagnostische aspect, is er in alle drie de toetsen gekozen voor drie niveaus van onderscheid; het overkoepelende vak als geheel (Nederlands, wiskunde of Engels), een aantal hoofdaspecten van dat vak, en een aantal deelaspecten binnen deze hoofdaspecten. De specifieke onderverdeling van deze niveaus verschilt per vak, en in sommige gevallen ook per jaar en per onderwijsniveau. Zoals ook zichtbaar in Tabel 1 zijn er voor Nederlands vier hoofdaspecten met elk drie deelaspecten (in één geval vier) en voor Engels vier hoofdaspecten met elk twee deelaspecten. Voor wiskunde zijn er twee hoofdaspecten met elk drie deelaspecten in 2016 en vijf hoofdaspecten met elk drie deelaspecten in 2017 (zie ook Tabel 2).[15] Één van de vijf wiskunde deelaspecten ('Informatieverwerking en onzekerheid') wordt alleen afgenomen bij havo- en vwo-leerlingen. De indeling heeft plaatsgevonden op basis van een panel van experts op elk vakgebied.

[15] In de DTT rapportages is uiteindelijk niet gerapporteerd op deelaspect voor wiskunde omdat er onduidelijkheid was over de validiteit van de gespecificeerde deelaspecten.

Zoals aangegeven in het vorige hoofdstuk heeft de indeling naar niveaus gevolgen voor de mate van precisie. De lagere opdeling op deelaspect voor Engels zou relatief gezien moeten leiden tot een hogere mate van precisie op dat niveau. Voor havo- en vwo-leerlingen op wiskunde zou verwacht mogen worden dat de opdeling naar vijf hoofdaspecten (in plaats van vier voor vmbo) tot lagere precisie op dat niveau leidt. Dit is echter geen wetmatigheid, aangezien verschillen in de kwaliteit van de vragen en de positie van de absolute grenswaardes er ook toe doen. Wanneer we de mate van precisie vergelijken tussen de verschillende vakken (door ofwel te kijken naar de hoogste van de drie PMP-waardes na de laatste vraag voor de categoriale score of naar de standaardfouten in de continue score), dan worden de verwachtingen op basis van de mate van uitsplitsing wel bevestigd. De Engels toets onderscheid zich inderdaad door een hogere precisie op vooral het deelaspect, ten opzichte van de Nederlands toets.[16] Deze verschillen zijn sterk, en illustreren daarmee dat een brede uitsplitsing sterke implicaties kan hebben voor de precisie van de diagnoses op deelaspect (zie ook de volgende sectie). Daarnaast blijkt ook dat na de toevoeging van het extra wiskunde-onderdeel voor havo-vwo leerlingen in 2017, de precisie op de andere wiskunde onderdelen inderdaad relatief sterker afneemt voor deze leerlingen, vergeleken met de vmbo-leerlingen.

Tot slot is een keuze gemaakt in het aantal categorieën dat wordt onderscheiden in het Latente Klassemodel. In de DTT is een keuze gemaakt voor drie categorieën. Dit is deels een conceptuele keuze geweest, die beantwoordde aan de behoefte om ook sterke punten van leerlingen te onderscheiden. De keuze voor het aantal categorieën heeft uiteraard ook consequenties voor de onzekerheid van de uitkomst. Hoe meer niveaus, hoe informatiever de uitkomst, maar ook hoe minder zeker we kunnen zijn dat dit ook echt de juiste categorie is voor die leerling (of school). Deze zekerheid is dus relatief het hoogst bij twee categorieën, maar die indeling is ook relatief het minst informatief. Aan het andere eind van het spectrum zit de continue indicator, waar we verder over uitwijden in het volgende hoofdstuk. De keuze die gemaakt is om van twee naar drie categorieën te gaan betekent dus automatisch dat de grenswaarde voor de ‘zekerheid’ minder vaak gehaald zal worden, wat vooral relevant is voor de deelaspecten.

3.5 Wat zijn de resultaten en hoe zijn deze verdeeld?

Één van de bepalende kenmerken van de DTT is de opsplitsing in deelaspecten, waarvoor sterke en zwakke punten geïdentificeerd kunnen worden. Vanuit dat perspectief is het zeer interessant om te bekijken in hoeverre deelaspecten binnen een bepaald hoofdaspect in de praktijk verschillen voor een bepaalde leerling. Scoren gemiddelde leerlingen op vrijwel elk onderdeel ook gemiddeld, of zien we juist vaker een combinatie van sterke en zwakke punten die zich in een gemiddeld eindoordeel vertaalt?

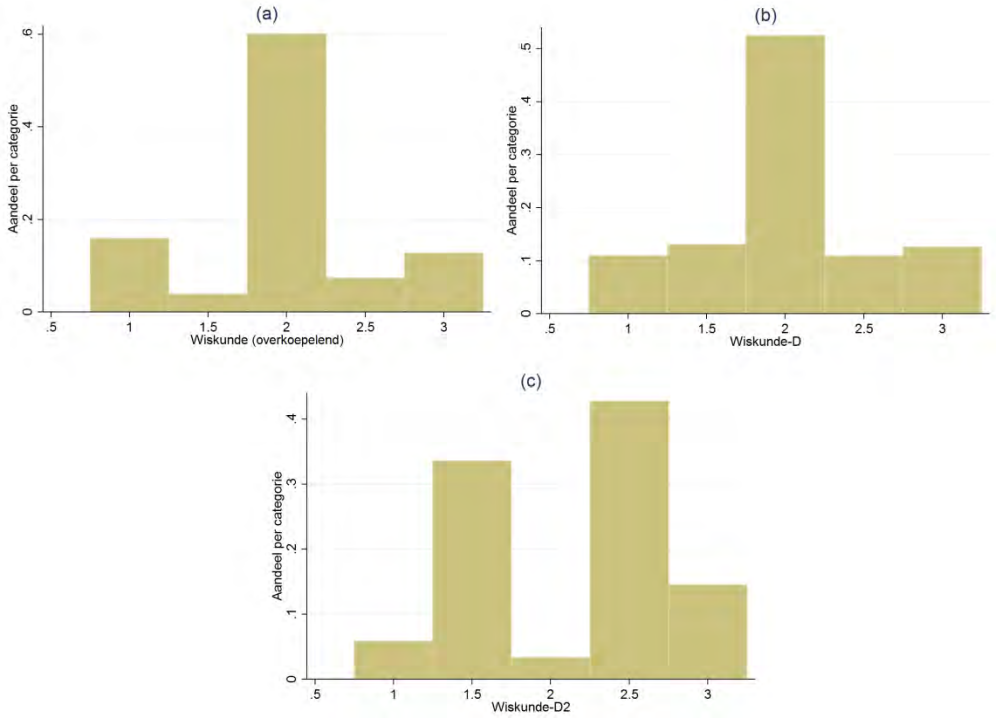
Om dit direct te kunnen analyseren kunnen we echter niet simpelweg kijken naar de variatie in de categoriale einduitkomsten, en wel om twee hoofdredenen. Allereerst kan variatie in die einduitkomsten ook gedreven worden door de verschillen in de posities van de absolute grenswaardes. Voor Engels zitten er bijvoorbeeld relatief veel meer leerlingen in de categorie ‘boven niveau’ en voor wiskunde juist meer in de categorie ‘onder niveau’. Maar ook binnen een vak zien we variatie in de verdelingen naar deelaspect. Zo scoren leerlingen binnen Nederlands bijvoorbeeld

[16] De Engels toets is ook op hoofdniveau preciezer, terwijl we daar aan de hand van de mate van uitsplitsing geen verschil zouden verwachten. Dit heeft waarschijnlijk te maken met het feit dat de gemiddelde kwaliteit per vraag bij Engels ook hoger ligt.

laag op Spelling en juist zeer hoog op Tekststructuur. Door deze variatie kan een leerling overal, ten opzichte van de andere leerlingen, in de buurt zitten van het gemiddelde maar dat kan zich dan op het ene onderdeel vertalen in 'onder niveau' en op een ander onderdeel in 'boven niveau'.

Een tweede punt is dat er veel onzekerheid zit rond de indeling van 'onder niveau', 'op niveau' en 'boven niveau'. De toets rapporteert na elke vraag de waarschijnlijkheid dat iemand tot één van elke categorieën behoort. Vaak is op het einde van de toets de hoogste van deze drie PMP-waardes nog steeds relatief laag. In Figuur 6 geven we deze onzekerheid aan, voor wiskunde als geheel, hoofdaspect Wiskunde D en deelaspect Wiskunde D2.[17] Wanneer de hoogste PMP-waarde onder de 0.9 valt, wijzen we de categorie 1.5 (als de onzekerheid tussen onder niveau of op niveau zit) of 2.5 (als de onzekerheid tussen op niveau of boven niveau zit) toe. De figuur laat zien dat de onzekerheid laag is voor wiskunde als geheel, omdat daar veel vragen zijn gesteld en er dus een hoge mate van precisie is. Bij de hoofdaspect en van wiskunde is de onzekerheid echter al vrij hoog (panel b) terwijl voor de deelaspecten bijna alle observaties onder de PMP-waardes van 0.9 vallen (panel c).[18]

Figuur 6: onzekerheid rond categorieën



Noot: Waarde van 1 verwijst naar 'onder niveau' (PMP_{onder}>0.90), waarde van 2 naar 'op niveau' (PMPop>.90) en waarde van 3 naar 'boven niveau' (PMP_{boven}>0.90). Voor de categorie 1.5 geldt: PMP_{op} <.90, PMP_{onder} <.90 en PMP_{onder} > PMP_{boven} en voor categorie 2.5 geldt PMP_{op} <.90, PMP_{boven} <.90 en PMP_{onder} < PMP_{boven}.

Op deelaspect weten we dus in het merendeel van de gevallen niet zeker in welke categorie een leerling valt. Aan de andere kant is een zekerheid van 90% wellicht een conservatieve aanpak.

[17] D verwijst naar het hoofdaspect 'Meten en Meetkunde', deelaspect 2 wijst daarbinnen naar 'Meerduidigheid'; zie Tabel 2).

[18] De resultaten zijn gebaseerd op de werkelijke leerlinggegevens van de DTT 2016 afname van wiskunde. Voor DTT 2017 ligt de onzekerheid nog iets hoger (met name voor domein D van wiskunde). Dit is een gevolg van het feit dat er meer hoofdaspecten van wiskunde zijn getoetst in 2017 waardoor het aantal vragen per onderdeel lager is.

Wanneer we bijvoorbeeld 80% zeker weten dat een leerling onder niveau is, dan lijkt dit ook nog steeds waardevolle informatie voor een leerling of school. Wanneer we bovenstaande oefening herhalen met een grenswaarde van 80%, dan neemt de onzekerheid inderdaad af, maar niet zeer sterk. Het aandeel onzekere observaties gaat van 38% naar 24% voor domein D van wiskunde en van 87% naar 76% voor subdomein D2.

Voor Nederlands zijn we grotendeels een vergelijkbaar beeld als voor wiskunde, al zijn de resultaten gemiddeld iets preciezer. De onzekerheid op hoofdaspect ligt daar gemiddeld op 35% en op deelaspect gemiddeld op 70%. De resultaten verschillen echter sterk voor Engels. Door zijn de aandelen observaties die de PMP-waarde van 0.9 niet halen 26% op hoofniveau en 41% op deelaspect. Zoals aangegeven in de vorige sectie komt dit deels door conceptuele verschillen in het meten van Nederlands versus Engels; de precisie is immers hoger op hoofdaspect terwijl er even veel domeinen zijn. Het verschil is echter veel groter op deelaspect, waar Engels maar twee domeinen heeft en Nederlands drie. Dit is indicatief bewijs dat een 'te verre' uitsplitsing sterke gevolgen kan hebben voor de nauwkeurigheid van de toets op deelaspect. Op leerlingniveau kunnen er daarom voor veel leerlingen geen sterke conclusies worden getrokken voor wiskunde en Nederlands, en is het sterk het overwegen waard om het aantal domeinen te reduceren (zeker aangezien er minder ruimte lijkt te liggen in het uitbreiden van de toetstijd).

Vanwege deze lage mate van precisie, is het dus onzeker of verschillen in de categoriale uitkomst komen door meetfout of door werkelijke verschillen in niveau. Het is daarom ook lastig om te zeggen in hoeverre vaardigheden op deelaspecten binnen een bepaald hoofdaspect in de pratijk sterk kunnen verschillen. Een alternatief is om te kijken naar een continue indicator (uitleg over de constructie van deze indicator volgt in Sectie 4.1). Ook deze indicator zal op de deelaspecten op leerlingniveau te maken hebben met significante meetfout, maar een nuttig aspect van de continue uitkomst is dat deze direct toelaat om direct te schatten of indicatoren statistisch significant van elkaar verschillen (op basis van de standaardfouten die door de IRT geschat worden). Wanneer we dit doen voor de deelaspecten van wiskunde, dan blijkt dat er in slechts 5% van de gevallen een dergelijk statistisch significant verschil is tussen de deelaspecten van een specifieke leerling. Die 5% is ook precies wat we zouden verwachten op basis van toeval (we werken in deze analyse immers met een 95% betrouwbaarheidsinterval). Dit illustreert een verdere complicatie op het deelaspect: aangezien er zo veel onnauwkeurigheid is over de uitkomstmaat, is bij het kleine aantal gevallen waarbij we (statistisch gezien) wel zeker genoeg lijken te zijn dat een leerling op twee onderdelen verschillend presteert, de kans relatief groot dat dit nu juist een 'toevalstreffer' is.[19] Dit is in de DTT vooral relevant voor wiskunde. Bij Nederlands hangt het sterk af van het hoofdaspect waarnaar we kijken hoe vaak deelaspecten van elkaar verschillen. Binnen 'Afstemmen op Doel en Publiek' en 'Teksstructuur' liggen deze aandelen tussen de 5% en 10%, voor 'Woord- en Zinsniveau' tussen de 10% en 15% en voor Spelling tussen de 15% en 20%. De spellingtoets kent relatief een sterke mate van precisie, omdat er veel (sub)vragen in zijn gesteld en ook vragen met een relatief sterk onderscheidend vermogen. Bij spelling kunnen we dus wel redelijk vaak met goede zekerheid zeggen dat leerlingen op onderdelen binnen spelling (spelling van werkwoorden, spelling van niet-

[19] Dit betreft hier de zekerheid over het feit dat twee deelaspecten werkelijk van elkaar verschillen, op basis van de continue indicator. Het gaat hier niet over hoe correct de diagnoses zijn in het Latente Klassemodel van de DTT.

werkwoorden en interpunctie) variëren in hun niveau, wat waardevolle diagnostische informatie is richting het formatieve leerproces.

Voor de overige onderdelen is het echter lastig in te schatten wat de meerwaarde is van het uitsplitsen naar deelaspecten. De meerwaarde van de deelaspecten is hoger, des te meer twee deelaspecten binnen hetzelfde hoofdaspect ook werkelijk verschillende vaardigheden vereisen. Wanneer leerlingen die slecht (goed) op D1 presteren ook altijd slecht (goed) presteren op D2, dan lijkt het onderscheid weinig informatief, aangezien ze aan dezelfde onderliggende vaardigheid lijken te relateren. Om na te gaan in hoeverre dit het geval is, kan ook naar correlaties gekeken worden. We schatten deze voor de geconstrueerde continue indicatoren. De correlaties tussen deelaspecten (binnen hetzelfde hoofdaspect) variëren redelijk sterk binnen en, vooral, tussen de vakken. Voor Nederlands liggen ze hoger dan voor wiskunde; rond de 0.50 gemiddeld met een uitschieter naar 0.65 voor de spellingsonderdelen. Correlaties voor de hoofdaspecten van Nederlands liggen tussen de 0.6 en 0.7. Voor wiskunde liggen de correlaties op deelaspect rond de 0.3, terwijl deze voor de hoofdaspecten rond de 0.6 liggen. Dit illustreert de dominante rol van meetfout[20] aangezien juist verwacht mag worden dat deelaspecten binnen, bijvoorbeeld, Meetkunde sterker aan elkaar zijn gerelateerd dan dat de hoofdaspecten Meetkunde en Verbanden/Formules aan elkaar zijn gerelateerd. De correlaties zijn het hoogst voor Engels; gemiddeld 0.78 voor de hoofdaspecten en 0.68 voor de deelaspecten.

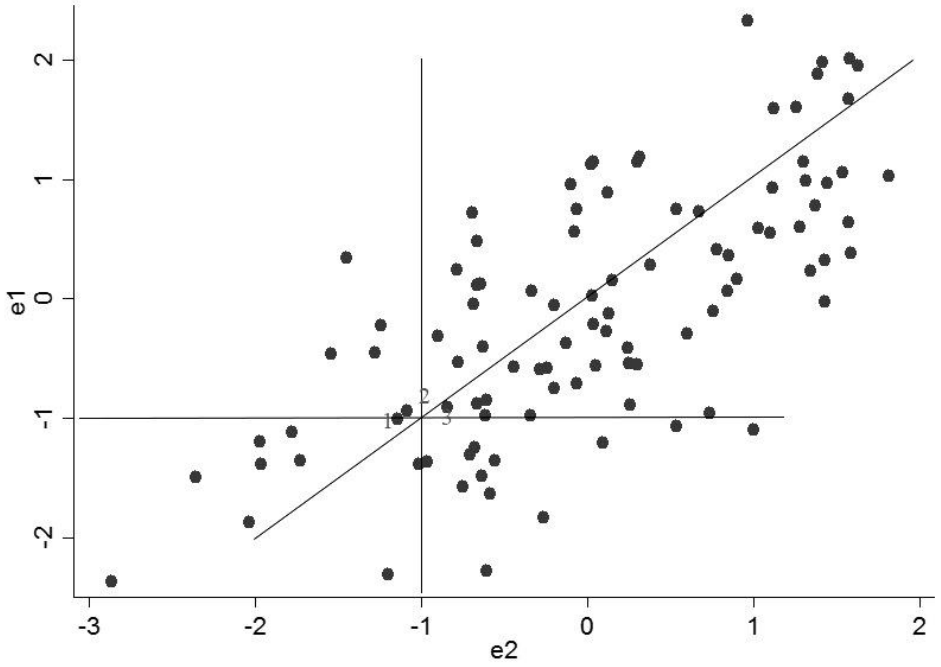
De correlaties tussen de deelaspecten zijn dus, zeker voor wiskunde en Nederlands, niet heel sterk, maar dit komt wederom voor een groot deel vanwege meetfout. Het is echter mogelijk om via IRT correlaties tussen variabelen te corrigeren voor meetfout, door beide indicatoren simultaan te schatten. Wanneer we dit doen, blijkt de gecorrigeerde correlatie tussen bijvoorbeeld deelaspect E1 en deelaspect E2 0.89 te bedragen.[21] De beide deelaspecten relateren dus heel sterk aan elkaar, wat suggereert dat het ware niveau van leerlingen op de beide onderdelen zeer dicht bij elkaar zal liggen. Anders gezegd, leerlingen die tekort komen op E1 zullen in het overgrote merendeel van de gevallen ook tekort komen op E2. Richting de formatieve functie van de toets zal het advies dan met name zijn om aan het onderdeel E (Verbanden en Formules) in zijn geheel meer aandacht te besteden.

De hoge meetfout voor de deelaspecten speelt zowel bij de categoriale indicator als ook bij de continue indicator. Welke uitkomstvariabele gebruikt wordt, heeft echter wel gevolgen voor de conclusies die getrokken kunnen worden. Dit wordt geïllustreerd in Figuur 7, waarin we continue indicatoren voor de wiskunde-onderdelen E1 en E2 tegen elkaar afzetten.[22] Als voorbeeld zetten we de scheidslijn tussen ‘onder niveau’ en ‘op niveau’ op -1, aangegeven door de horizontale en verticale scheidslijnen in de figuur. Drie leerlingen zijn hierbij uitgelicht. Leerling 1 is onder niveau voor beide onderdelen, leerling 2 is op niveau voor E1 en onder niveau voor E2 en leerling 3 is op niveau voor beiden. Alle drie deze leerlingen hebben op beide onderdelen een vrijwel identieke schatting, die zeker in het licht van de meetfout niet van elkaar te onderscheiden is. De categoriale

[20] Het gaat hier specifiek over de meetfout binnen het geschatte IRT 2PL model maar de hoge meetfout/lage mate van precisie speelt op het niveau van de deelaspecten bij alle schattingsmethodes.
[21] Tussen de andere deelaspecten van wiskunde en ook tussen de deelaspecten van Nederlands vinden we vergelijkbare gecorrigeerde correlaties.
[22] Het betreft hier een simulatie van waardes voor E1 en E2, gebaseerd op de geschatte correlatie tussen deze twee deelaspecten in de 2017 afname van de DTT.

indicator leidt echter tot drie zeer verschillende conclusies. Er kan dan gekozen worden om geen uitkomst te rapporteren,[23] maar het feit dat een leerling op de grens zit van onder en op niveau is ook interessante informatie. Als een tussenoplossing kan er met de categoriale indicator gekozen worden om ook die onzekerheid in de rapportage weer te geven. Dit is in feite wat we gedaan hebben in Figuur 6. Hiermee wordt er echter in feite een nieuwe indicator met vijf categorieën gecreëerd. Bovendien is het ook met die indicator informatief om te weten hoe dicht iemand bij die grenzen zit. De categoriale indicator kan dan wel steeds uitgebreid worden om dit weer te geven, maar als dit doorgevoerd wordt eindigen we uiteindelijk weer bij een continue indicator, die daarmee dus in feite meer informatie geeft.[24]

Figuur 7: continue indicatoren voor E1 en E2 (wiskunde)



Bovenstaande discussie geldt voor de individuele scores. Op schoolniveau is de meetfout van de schattingen veel lager (zie ook het volgende hoofdstuk), omdat scores van meerdere leerlingen geaggregeerd kunnen worden. Gezien de kleine meetfout op schoolniveau zijn in dit geval de ruwe correlaties op schoolniveau zeer informatief over de mate van samenhang van deze schoolscores. Wanneer we deze correlaties bekijken voor de verschillende deelaspecten binnen een bepaald hoofdaspect, dan zijn er wederom verschillen naar vak.[25] De gemiddelde correlaties liggen rond de 0.9 voor Engels, rond de 0.75 voor Nederlands, en rond de 0.45 voor wiskunde. Voor Engels geldt dus dat scholen die goed scoren op, bijvoorbeeld, domein 3.1 ook bijna altijd goed scoren op domein

[23] In de DTT wordt dit niet gedaan voor leerlingen op de grens van de scheidslijn, maar voor leerlingen met onvoldoende zekerheid (in feite wordt niet gerapporteerd wanneer het aantal onjuiste/inaccurate rapportages in de populatie heel hoog ligt). Hoewel deze zaken gerelateerd zijn, zijn ze niet hetzelfde.

[24] De rapportages van de DTT pilot gaan in zekere zin verder dan drie categorieën, omdat met de grootte van de symbolen ook het niveau van zekerheid wordt weergegeven. Een laag niveau van zekerheid betekent echter niet automatisch dat iemand dicht bij de grens zit (alleen dat het onduidelijk is aan welke kans van de grens hij of zij valt).

[25] De scores worden hier geaggregeerd binnen elk onderwijsniveau op elke school.

3.2. De opsplitsing in deelaspecten is in dat geval dus minder informatief. Voor Nederlands en met name wiskunde zijn er aan de andere kant wel relatief veel gevallen waarin de scholen verschillend scoren op de verschillende deelaspecten binnen een hoofdaspect.

3.6 De kijk van scholen op de indeling van de DTT in hoofdaspecten en deelaspecten

Aan twaalf Limburgse scholen zijn de leerlingmodellen van de DTT Nederlands en de DTT wiskunde voorgelegd. Bij het onderzoeken van de gewenste breedte en diepte van de te toetsen vaardigheden is volgens de scholen een juiste balans tussen algemeen en specifiek van belang. Bij een te algemeen niveau van rapporteren zal de toets onvoldoende bruikbaar zijn om het leren verder te brengen. Een verregaande uitsplitsing binnen de hoofdonderdelen brengt echter het risico tot beperkte bruikbaarheid met zich mee omdat docenten en leerlingen niet goed meer weten waar te beginnen.

Bij het leerlingmodel van de DTT Nederlands geven de meeste scholen aan dat de huidige indeling in hoofd- en deelaspecten goed werkbaar is voor het gebruik van het instrument op leerlingniveau. Een enkeling geeft aan dat eigenlijk nog meer diepte gewenst is. Een rapportage op alleen de hoofdaspecten zou in elk geval te algemeen zijn. Verder geven alle scholen aan dat het cruciaal is dat het voor docent (en leerling) duidelijk is welke vervolgstappen gezet kunnen worden als gevolg van de uitslag op de DTT. Voor de scholen is het van groot belang dat dit op eenvoudige wijze automatisch uit de rapportage naar voren komt.

Over het leerlingmodel DTT wiskunde is wat meer verdeeldheid. Sommigen geven aan dat een rapportage op domeinniveaus (hoofdaspecten) voldoende is, aangezien de meeste methodes een zelfde indeling hanteren en de DTT daar mooi op aansluit. Anderen geven aan dat een rapportage op alleen domeinniveaus weinig toegevoegde waarde heeft. Docenten weten vaak al goed hoe hun leerlingen het doen op de domeinen. Bovendien komt dit type informatie ook naar voren uit andere instrumenten, waarvan scholen gebruik maken. Ook voor de DTT wiskunde geldt, net als bij de DTT Nederlands, dat scholen vooral behoefte hebben aan directe vervolgstappen die passen bij de verschillende onderdelen en mogelijke toetsuitslagen.

Conclusie

In dit hoofdstuk is gekeken naar keuzes die zijn gemaakt in de opzet van de DTT pilot, wat betreft bijvoorbeeld de toetsonderdelen, items en inpassing van adaptiviteit. Onze simulaties laten zien dat de meerwaarde van adaptieve afname van vragen in de DTT verder vergroot zou kunnen worden, wanneer er niet alleen gekeken wordt naar hoe vaak een diagnose gesteld kan worden, maar ook naar hoe snel dat punt bereikt wordt.

Een ander belangrijk punt in de opzet van diagnostische toetsen is in hoeverre deelaspecten binnen een bepaald vak (bijvoorbeeld deelaspecten van meetkunde binnen wiskunde) daadwerkelijk van een verschillend niveau zijn voor een bepaalde leerling. Hoe vaker dit het geval is, hoe groter de meerwaarde van een dergelijke uitsplitsing in de toets. Aan de andere kant, wanneer leerlingen die tekort komen op meetkunde ook consistent tekort komen op alle deelaspecten van meetkunde, dan is deze opsplitsing minder waardevol. De resultaten laten zien dat de nauwkeurigheid op dit niveau van de toets in het algemeen te laag is om te kunnen bepalen of deze deelaspecten wel of niet van elkaar verschillen, maar dat ook de correlatie tussen deze onderdelen als ze precies zouden worden

gemeten vrij hoog is. Door scholen wordt aangegeven dat de informatie op het niveau van de deelaspecten juist vooral waardevol is voor het formatieve leerproces. Voor de waarde van de toets op individueel niveau lijkt daarom een minder brede uitsplitsing van deelaspecten (zoals bijvoorbeeld in de Engels toets van de DTT) wenselijk, ook gezien de benodigde toetstijd door scholen al als erg ruim werd ervaren. Voor de waarde van de toets op schoolniveau spelen deze afwegingen veel minder, aangezien de precisie van de scores daar veel hoger ligt.

4. Alternatieve scoringsmogelijkheden

In de in de pilot ontwikkelde opzet van de DTT wordt er een indicatie van onder niveau, op niveau of boven niveau gegeven voor leerlingen. Een alternatieve optie is om een continue indicator te gebruiken. Een continue indicator geeft aan hoe ver een leerling eventueel boven of onder niveau presteert op een bepaald onderdeel. Voor deelaspecten zal een dergelijke aanpak aanlopen tegen de beperkte nauwkeurigheid van de toets. Op schoolniveau kan echter een hogere nauwkeurigheid worden behaald en kan een continue indicator meer informatie geven.In dit hoofdstuk construeren we een dergelijke indicator door gebruik te maken van een IRT 2PL model (zie uitleg in Sectie 2.1). We analyseren de precisie van deze indicator op zowel individueel niveaus als schoolniveau, en bekijken hoe deze zich verhoudt tot de categoriale uitkomst die berekend wordt in de DTT.

4.1 Een continue indicator

Op basis van de afnamedata van de DTT pilot construeren we via een 2pl IRT model continue vaardigheidsscores, op de verschillende aspecten van de DTT. Het gebruik van deze continue indicator heeft een aantal belangrijke voordelen. De belangrijkste reden is dat het toelaat om meer precieze schoolscores te construeren, zoals aangetoond zal worden in dit hoofdstuk. Als we kijken naar de uitkomsten op leerlingniveaum dan heeft de continue indicator als nadeel dat deze vanuit psychometrisch oogpunt restrictiever is dan bijvoorbeeld het latente klassemodel.[26] Aan de andere kant heeft de continue aanpak ook op leerlingniveau een aantal voordelen.[27] Deze voordelen ten opzichte van het klassemodel hebben allemaal te maken met het feit dat de feitelijke verdeling van de onderliggende vaardigheden die we in elk model meten wel continue is. In de eerste plaats blijkt dat er voor leerlingen die heel dicht bij de grens zitten tussen twee klassen er veel vragen nodig zijn om een hoge PMP te halen.[28]In de tweede plaats roept de indeling in drie klassen de vraag op bij iemand die bijvoorbeeld onder niveau scoort, of hij net onder niveau of ver onder niveau zit. Als een school besluit om deze leerling extra aan deze vaardigheid te laten werken is dat belangrijke informatie. Ten derde ontstaat door deze aanpak een vrij arbitraire grens tussen de drie klassen. Terwijl bij de normering van de toets een duidelijke grenswaarde is bepaald die een leerling minstens zou moeten halen, wordt bij de uitvoering van de toets bekeken of het antwoordpatroon van een leerling meer lijkt op de gemiddelde leerling die onder niveau heeft gescoord of de gemiddelde leerling die op niveau heeft gescoord. De gemiddeldes van deze twee groepen kunnen echter heel anders zijn dan de antwoordpatronen van leerlingen die dicht tegen de grens aanzitten.

We benadrukken verder dat een continue score niet betekent dat de toets summatief wordt. Een continue score impliceert een bepaalde rangschikking, maar dat geldt net zo goed voor een indicator met drie categorieën. Voor de eerste zijn er simpelweg meer mogelijke uitkomsten. Ook met een continue score kunnen diagnoses gesteld worden over waar leerlingen of scholen achterblijven of uitblinken, die vervolgens voor formatieve doeleinden kunnen worden gebruikt door zowel leerling, leraar als school.

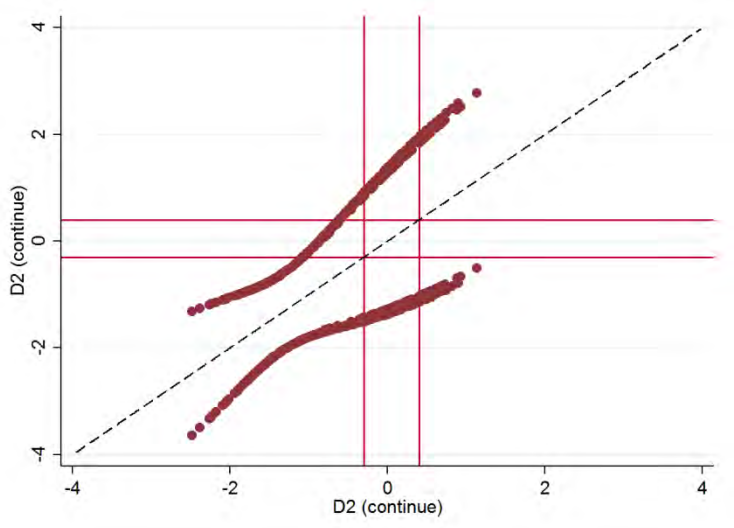
[26] Een voorbeeld is dat het IRT model oplegt dat het verschil in de kans op het goed beantwoorden van een vraag symmetrisch verdeeld is over de distributie.
[27] Een deel van deze voordelen is aan bod gekomen in Sectie 3.5.
[28] De convergentie is dan langzaam, omdat de methode leerlingen wil indelen op basis van een gemiddeld profiel in elke categorie, terwijl de antwoordpatronen van deze 'grensleerlingen' verschillen van elk van deze gemiddeldes.

We bekijken in deze sectie concreet, op basis van de data van de DTT pilot, wat de precisie van de continue schattingen is op zowel individueel niveau als schoolniveau. We analyseren eerst het individuele niveau. Aangezien het aantal vragen per deelaspect beperkt is, zal de continue indicator op dit niveau van de toets een relatief lage precisie hebben. Voor de hoofdaspecten en voor het toetsonderwerp als geheel zal dit veel minder een probleem zijn.

Als voorbeeld nemen we een deelaspect van wiskunde, gelabeld in de DTT als 'D2'. In Figuur 8 zien we de precisie van deze deelscore op individueel niveau. De figuur laat de 95% betrouwbaarheidsintervallen zien. Voor elke score op de 45 graden lijn zien we de bandbreedte waarvoor met 95% zekerheid te zeggen is dat het werkelijke niveau van de leerling hier binnen ligt. De gemiddelde score is op 0 gezet, en de standaardafwijking bedraagt 1. De figuur laat zien dat er hogere precisie is bij lagere scores; daar is de bandbreedte smaller. Dit geeft aan dat de relatief gemakkelijke vragen beter onderscheidend zijn dan de relatief moeilijke vragen (en/of dat er meer makkelijke vragen gesteld zijn voor dit onderdeel).

Tegelijkertijd is de precisie gemiddeld laag. Voor iemand met een geschat gemiddeld niveau is er een redelijke kans dat hij of zij een standaardafwijking boven of onder dat gemiddelde zit. In dit geval is het daarom redelijk om categoriale indicator te gebruiken, aangezien we voor de gemiddelde inschatting alleen zeker weten dat diegene 'ongeveer' rond het gemiddelde zit. Aan de andere kant bleek uit eerdere analyses dat er ook bij de categoriale indicator vaak nog veel onzekerheid is (zie Sectie 3.5). Dat zien we ook in onderstaande figuur; ook bij de allerhoogste scores is er nog steeds een redelijke kans dat diegene een gemiddeld niveau heeft (de score 0 valt nog binnen de betrouwbaarheidsinterval). Bij lage scores hebben we onder een score van -1 genoeg zekerheid om te zeggen dat deze leerlingen onder niveau zijn.

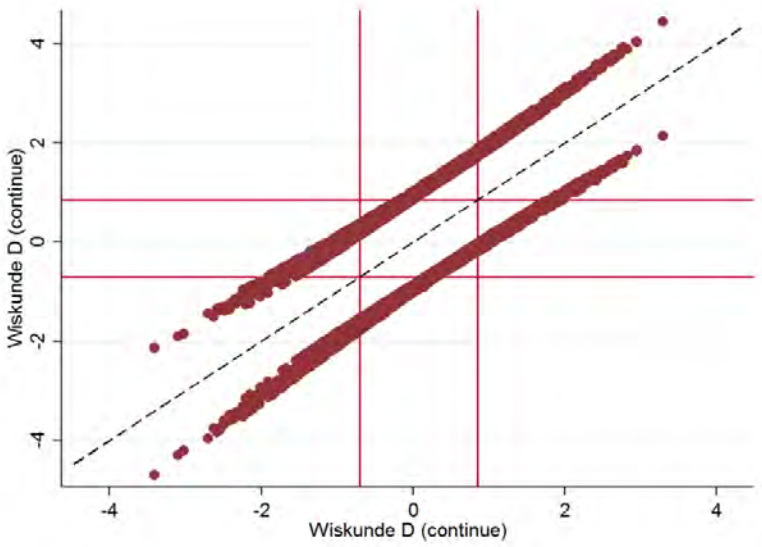
Figuur 8: betrouwbaarheidsinterval wiskunde D2



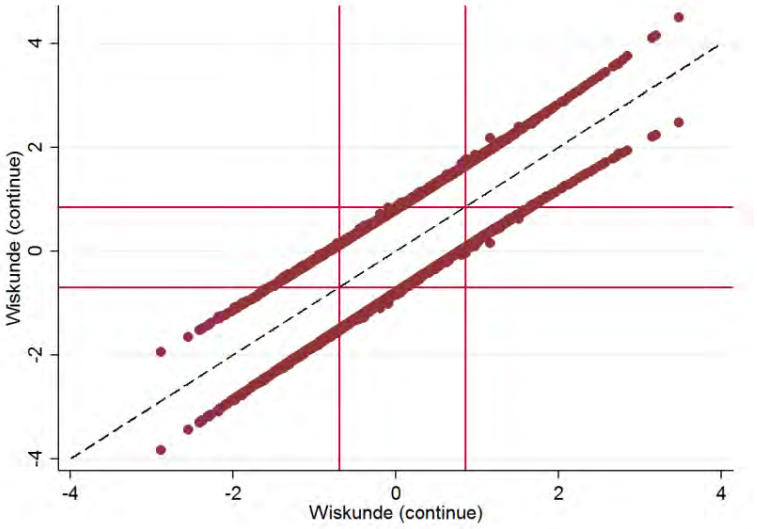
Als we kijken naar de hoofdaspecten, zoals meten en meetkunde (onderdeel D), zal de onnauwkeurigheid automatisch lager zijn, omdat er meer vragen beschikbaar zijn om de inschatting op te baseren. Figuren 9 en 10 laten zien de schattingen voor het hoofdaspect D van wiskunde en voor wiskunde als geheel veel preciezer zijn. Waar de betrouwbaarheidsinterval van D2 bij de gemiddelde leerling loopt van -1.3 tot 1.3, is dit voor D gelijk aan -0.95 tot 0.95 en voor wiskunde als

geheel gelijk aan -0.75 tot 0.75. Deze toename in precisie is belangrijk voor de afweging tussen een categoriale en een continue indicator. Hoe meer precisie, hoe meer informatie er verloren gaat wanneer we overschakelen naar een categoriale indicator (waarin immers iedereen binnen hetzelfde niveau gelijk gesteld wordt).

Figuur 9: betrouwbaarheidsinterval wiskunde D



Figuur 10: betrouwbaarheidsinterval wiskunde totaal

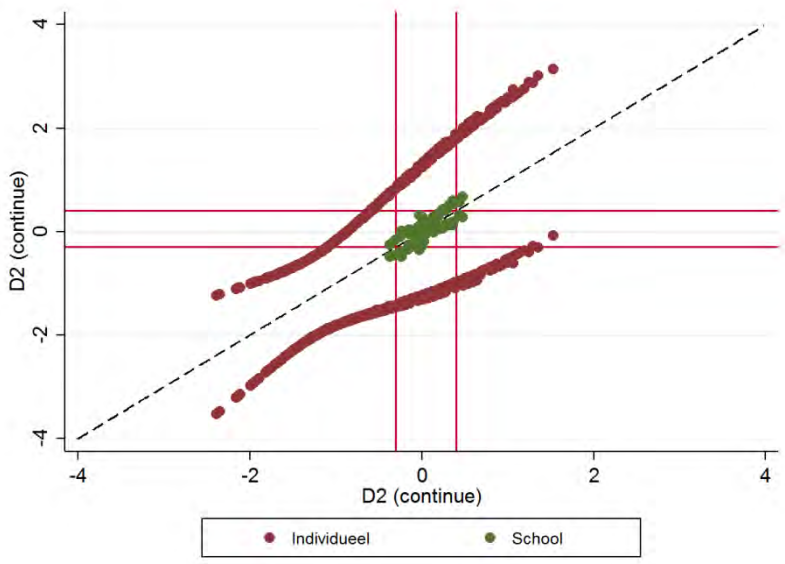


4.2 De continue schoolscore

De mate van precisie is dus redelijk zwak op het niveau van de deelaspecten, zoals D2. De situatie in Figuur 8 geldt echter op individueel niveau. Op schoolniveau is er een veelvoud aan informatie aanwezig, aangezien we voor een hele groep leerlingen informatie per vraag hebben. Op dezelfde manier dat de precisie toeneemt in de Figuren 1 naar 3 naarmate er meer informatie wordt toegevoegd, geldt dit ook als we overstappen van individueel niveau naar schoolniveau. Dit is zichbaar in Figuur 11, waar nu ook het betrouwbaarheidsinterval voor de schoolscore is toegevoegd.

De schoolscore is hier simpelweg het gemiddelde van alle leerlingscores per school.[29] De schoolscore is zeer precies.

Figuur 11: betrouwbaarheidsinterval individueel en school (D2)

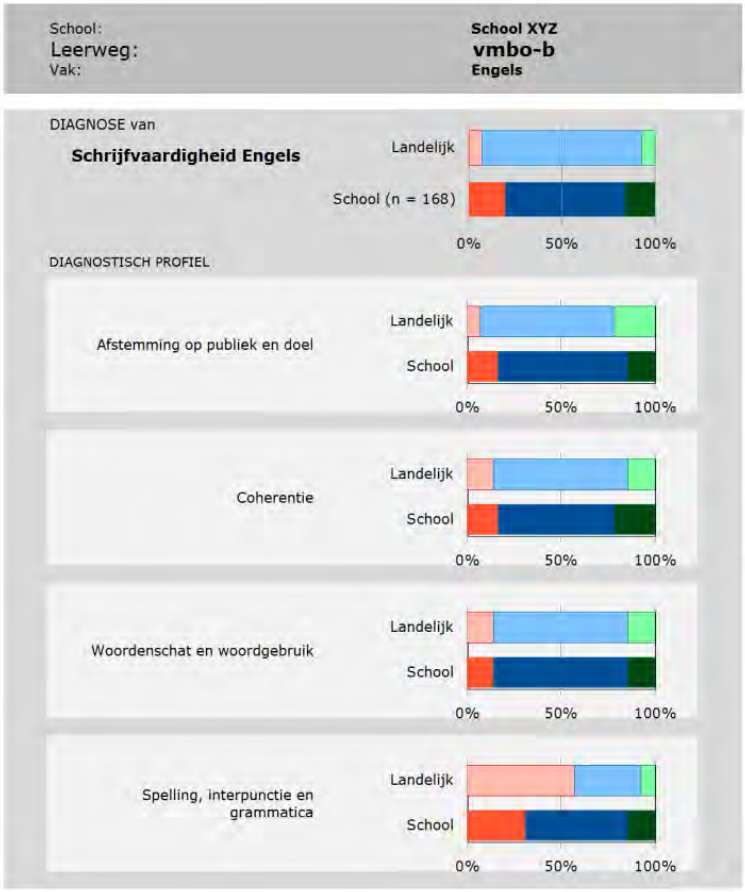


4.3 Hoe verhoudt de continue indicator zich tot de categoriale?

Op basis van de informatie die direct aanwezig is in de DTT kunnen prestaties op schoolniveau bekeken worden door de categoriale score op individueel niveau samen te voegen. Hieruit kan bijvoorbeeld het gemiddelde van de drie uitkomsten worden genomen, of het aandeel leerlingen dat onder dan wel boven niveau zit. In de DTT-rapportage tijdens de pilot werd aan scholen de aandelen leerlingen voor elke categorie getoond, vergeleken met het beeld voor alle leerlingen die de test hebben afgenomen. Een voorbeeld hiervan is zichtbaar in Figuur 12.

[29] Scores zijn feitelijk berekend per onderwijsniveau van elke school, waarbij het gemiddelde van alle leerlingen van elk onderwijsniveau op 0 is gezet

Figuur 12: voorbeeld huidige schoolrapportage DTT



Een alternatief is om de schoolrapportage te baseren op de schoolscores gebaseerd op de IRT. Een eerste vraag is hoe deze continue schoolscore zich verhoudt tot de samengestelde schoolscore in de DTT pilot. Bovenstaand voorbeeld geeft indirect aan wat hier de beperking van is. Voor de schrijfvaardigheid van Engels als geheel zijn de twee uiterste categorieën zeer klein. Binnen de blauwe middengroep wordt iedereen gelijkgesteld, terwijl er binnen die groep grote verschillen zullen zijn. Aangezien er zeker voor het overkoepelende vak als geheel en ook voor de hoofdaspecten veel vragen zijn om het niveau precies te schatten, gaat hier dus veel informatie in verloren. De school in het voorbeeld ziet dat er bij hen wat meer leerlingen zijn aan beide extremen, maar hebben weinig indruk van hun gemiddelde niveau, terwijl dat in potentie sterk kan afwijken van de rest van de scholen.

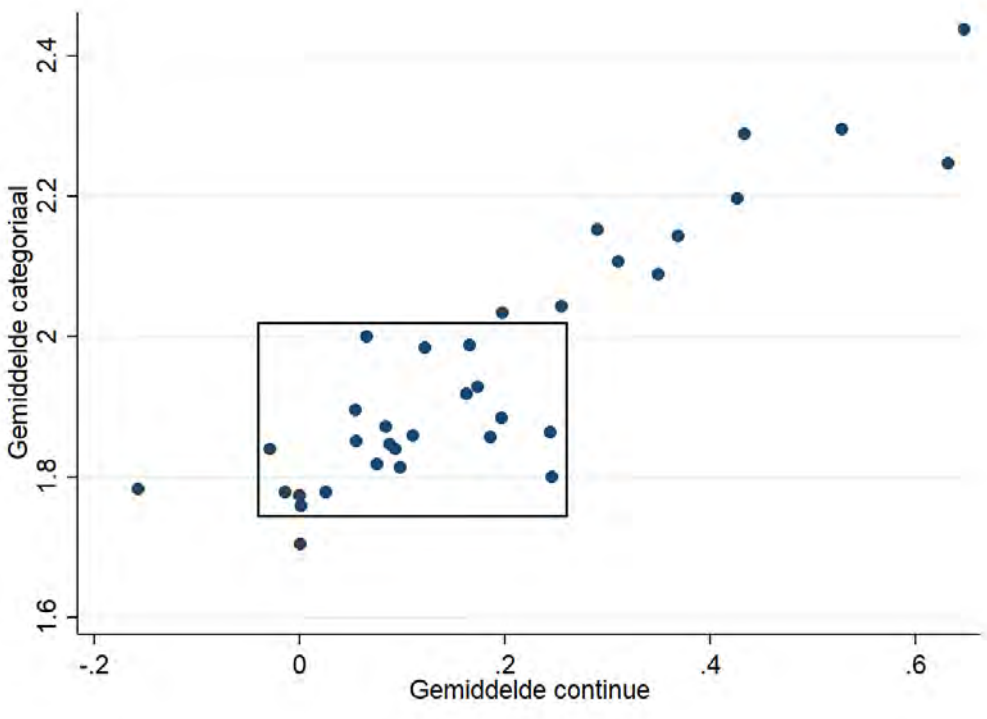
Met de continue indicator kan er dus een gemiddelde score op schoolniveau worden samengesteld die zeer precies is. Wanneer een school ook een indruk wil hebben van de aandelen achterblijvende en uitblinkende leerlingen, dan kunnen er binnen de continue score ook bepaalde grenswaarden worden gekozen, en eenzelfde beeld als in Figuur 12 worden weergegeven. Maar de meerwaarde van de continue indicator ligt dus in het leveren van meer precieze gemiddelde scores.

Om een concreter beeld te krijgen van de meerwaarde van de continue indicator, vergelijken we deze met de samengevoegde categoriale uitkomsten op schoolniveau. Op basis van de categoriale uitkomst kunnen we de aandelen onder niveau, op niveau en boven niveau gebruiken als vergelijking, of we kunnen kijken naar het gemiddelde van de drie categoriale scores (waarbij onder niveau telt voor 1, op niveau voor 2 en boven niveau voor 3).

In Figuur 13 vergelijken we de gemiddelde continue score met de genoemde gemiddelde categoriale score (voor VWO scholen alleen). We kiezen hiervoor het wiskunde deelaspect D1, omdat deze een symmetrische verdeling heeft van de categorieën. Dat wil zeggen, het aandeel scores onder niveau is vrijwel identiek aan het aandeel scores boven niveau. Mocht bijvoorbeeld het aandeel onder niveau veel kleiner zijn dan het aandeel boven niveau, dan zou dat een reden kunnen zijn om de onder niveaus scores zwaarder te laten wegen, wat niet gebeurt wanneer we simpelweg het gemiddelde nemen. Door hier een deelscore te gebruiken die symmetrisch verdeeld is, is de vergelijking dus meer redelijk.

Figuur 13 toont aan dat er een duidelijke positieve relatie is tussen beide scores, wat ook verwacht mag worden. Maar binnen het rechthoekige 'middenveld' is die correlatie zeer zwak tot miniem. De twee type scores geven dus een zeer vergelijkbaar beeld voor de 'extreme' scholen, maar kunnen sterk verschillen voor de scholen daar tussen in. De scholen die linksonder in het vierkant zitten zijn volgens de categoriale indicator even goed als de scholen die rechtsonder zitten, maar volgens de 'rijkere' continue indicator zitten deze ruim 0.2SD in score uit elkaar. Anders gezegd, de scholen rechtsonder horen volgens de categoriale indicator tot de allerzwaksten, terwijl ze in feite duidelijk boven de mediane school zitten. Andersom bekeken zitten er in het midden van de rechthoek zeer vergelijkbaar presterende scholen, die volgens de categoriale indicator redelijk sterk verschillen. Het feit dat de verschillen sterk zijn voor de middengroep is niet verrassend; daar zitten vooral veel leerlingen die in de brede categorie 'op niveau' vallen, en waarbij er dus relatief veel informatie verloren gaat door ze allemaal gelijk te stellen binnen die categorie.

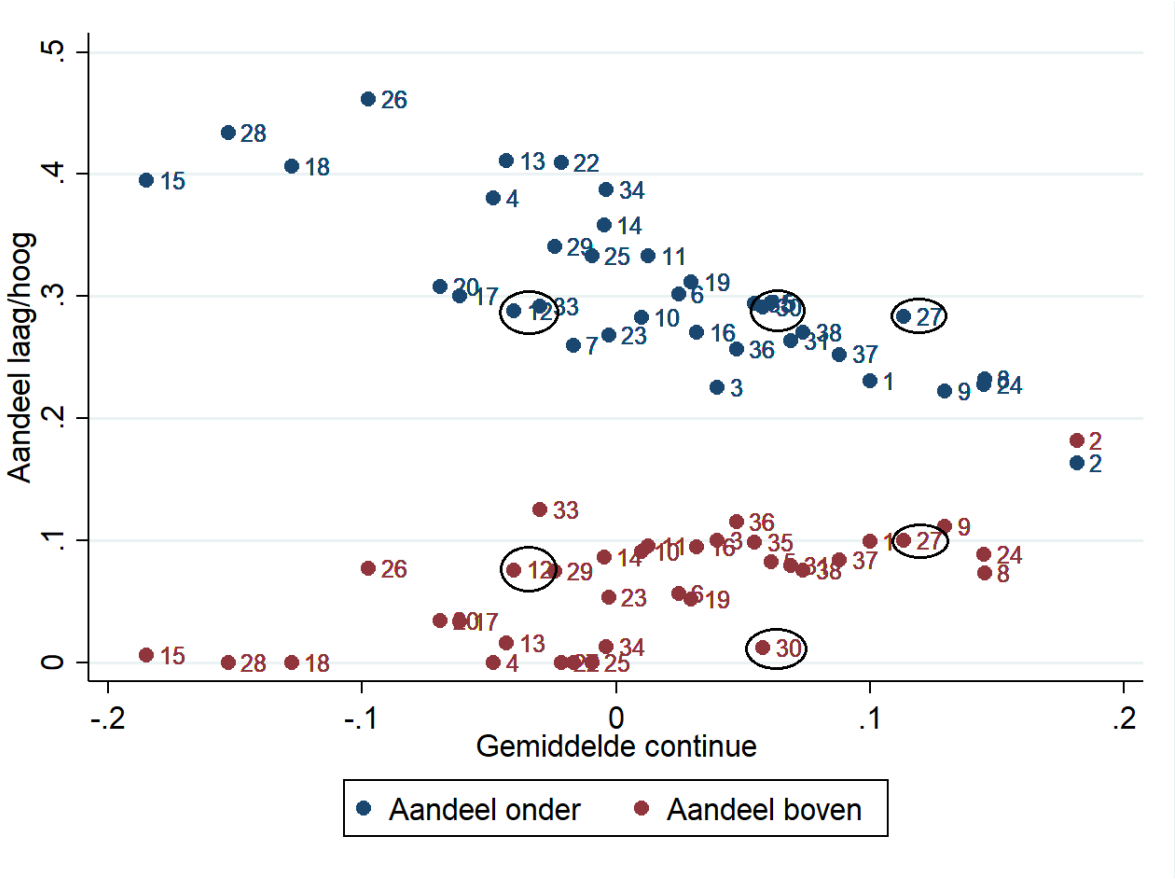
Figuur 13: continue score vs. categoriale score (schoolniveau)



In Figuur 14 gebruiken we voor de categoriale indicator de hele verdeling aan scores, in plaats van een geconstrueerd gemiddelde. De figuur vergelijkt de gemiddelde continue scores met zowel de aandelen onder niveau als de aandelen boven niveau. Scholen met een hogere gemiddelde continue

score hebben logischerwijs meer leerlingen boven niveau en minder leerlingen onder niveau. We zien alleen ook dat zelfs scholen met vrijwel identieke aandelen boven en onder niveau (en dus automatisch ook vrijwel identieke aandelen op niveau) nog steeds sterk kunnen verschillen in de continue score. De scholen 12 en 27 zijn hier een duidelijk voorbeeld van. Daarnaast geldt voor school 30 dat ze vergeleken met school 12 even veen leerlingen onder en *minder* leerlingen boven niveau hebben, maar toch een duidelijk hogere continue score hebben.

Figuur 14: continue score vs. aandelen laag/hoog



De verschillen in bovenstaande figuren zijn niet verwaarloosbaar. Een verschil van 0.2SD betekent het verschil tussen een slecht presterende school en een gemiddeld presterende school, en bedraagt ook meer dan de helft van het verschil tussen de gemiddelde havo-school en de gemiddelde vwo-school. Dit niveau van precisie is vooral belangrijk wanneer diagnostische toetsen uitgebreid worden naar meerdere afnames om te kijken of het niveau van een leerling of een school verbeterd is. Zelfs een school die sterk verbetert op een bepaald onderdeel zal niet meteen van de slechtst presterende school naar de best presterende school schieten. Een sprong voorwaarts van 0.2SD zal al een zeer opzienbarende verbetering zijn. Het is daarom van cruciaal belang dat we zulke verbeteringen zo precies mogelijk kunnen identificeren.

Een meer formele wijze om te kijken in hoeverre het uitmaakt of we naar categoriale scores of naar continue scores kijken is om naar de correlaties te kijken tussen de twee variabelen. De correlatie voor het vak als geheel (dus bijvoorbeeld wiskunde) ligt rond de 0.75; de correlatie voor de hoofdaspecten ligt tussen de 0.65 en 0.80 en de correlatie voor de deelaspecten ligt tussen de 0.50 en 0.75. Gezien het feit dat beide scores op exact dezelfde set vragen zijn bepaald, zijn dit relatief

lage correlaties, zeker voor de deelscores. Hieruit blijkt dus wederom dat het wel degelijk uitmaakt of we de continue score gebruiken of de categoriale.

4.4 Absolute of relatieve grenswaardes voor categorieën?

In de DTT pilot is gekozen voor 'absolute' standaarden die van tevoren zijn bepaald. De continue score die in dit onderzoek is samengesteld is per definitie relatief. Maar ook binnen een categoriale indicator kan ervoor worden gekozen om met relatieve scheidslijnen te werken. Daarbij zijn er verschillende mogelijkheden. Een traditionele aanpak zou zijn om leerlingen steeds te positioneren op basis van de prestaties van alle andere leerlingen binnen hun onderwijsniveau. Daarbij zou dus bijvoorbeeld de onderste 33% de laagste categorie toegewezen en de hoogste 33% de hoogste categorie. In dat geval worden de grenswaardes dus ook gelijk getrokken tussen de verschillende deelaspecten en is duidelijker zichtbaar wanneer een leerling achter ligt of voor ligt ten opzichte van leerlingen op hetzelfde onderwijsniveau.

Een andere mogelijkheid is om steeds het niveau van de leerling op het hoofdaspect als uitgangspunt te gebruiken, en dan te kijken wat daarbinnen sterke of zwakke punten zijn. Bij een leerling die bijvoorbeeld hoog scoort op wiskunde kunnen dan onderdelen die vanuit absoluut oogpunt gemiddeld zijn aangemerkt, als relatief zwakke punten gelden, gegeven het algemene niveau van de leerling op wiskunde.

Voordelen van een absolute grenswaarde is dat de norm onafhankelijk is van de prestatie van de leerlingen op de toets en van de toetsinhoud. Een dergelijke grenswaarde zal vooral waarde hebben on relatie tot de vraag wat leerlingen zouden moeten kunnen (wat in het geval van de DTT gesteld is door SLO), en biedt daarbij voor leerlingen en scholen ook goede inzichten richting een centraal eindexamen.

Een mogelijk nadeel van absolute grenswaardes is dat ze altijd deels subjectief zijn. Ze worden bepaald door experts, die het ook onderling niet volledig eens zullen zijn en een compromis sluiten. Dat compromis zal ook niet volledig los te zien zijn van hoe leerlingen in de realiteit presteren. De bepaling van wat een leerling van een bepaalde leeftijd 'zou moeten kunnen' zal immers altijd samenhangen met wat leerlingen in de praktijk blijken aan te kunnen. Bovendien is de standaard een gemiddelde afweging van alle type leerlingen binnen dat niveau. Voor scholen die bijvoorbeeld een ander curriculum hebben zijn de standaarden daarom wellicht niet altijd van toepassing. Het voordeel van de relatieve grenswaarde is dat duidelijk is wat de standaard is en hoe deze geïnterpreteerd moet worden (namelijk niveau ten opzichte van de prestaties van de rest van de steekproef), terwijl dit bij een absolute grenswaarde abstracter is. Daarnaast kunnen relatieve grenswaardes ook verder uitgebouwd worden door de grenswaarde bijvoorbeeld op het gemiddelde niveau van de school of op het gemiddelde niveau van de leerling voor het onderdeel als geheel te plaatsen. Absolute grenswaardes laten die flexibiliteit niet toe. Verder is het ook praktisch makkelijker en sneller om een relatieve grenswaarde te bepalen (wat werkt met een druk op de knop) dan om elke keer het proces door te lopen met experts om absolute grenswaardes te bepalen.

Zeker wanneer diagnostische instrumenten herhaaldelijk worden afgenomen is dat een belangrijke overweging.[30]

4.5 De kijk van scholen op de alternatieve scoringsmogelijkheden

Aan twaalf Limburgse scholen is de vraag voorgelegd welke score hun voorkeur heeft in de leerlingrapportage: een continue of een categoriale score. Hierbij heeft men de oorspronkelijke DTT rapportage (met categorieën onder, op en boven niveau) kunnen vergelijken met de alternatieve, hierboven besproken leerlingrapportage. Aan de scholen zijn hierbij de scores niet uitgedrukt in standaard deviaties zoals in de figuren in dit hoofdstuk, maar in percentielscores lopend van 1 tot 100, aangezien deze voor scholen meer intuïtief zullen zijn. Bijna alle scholen geven de voorkeur aan het gebruik van deze percentielscores boven de categoriale indicator uit de DTT, omdat deze de relatieve scores van de leerling preciezer weergeven. Bovendien geven de percentielscores direct een indruk van de verdeling van de diagnose op de verschillende deelaspecten van de vaardigheid.

Ook voor de schoolrapportage is aan scholen de vraag voorgelegd welke score hun voorkeur heeft: een continue of een categoriale score. Hierbij heeft men de oorspronkelijke DTT rapportage (met categorieën onder, op en boven niveau) kunnen vergelijken met de alternatieve, hierboven besproken, schoolrapportage (met percentielscores). Sommige gesprekspartners geven aan dat de kleuren in de DTT rapportage prettig zijn, omdat deze meteen een indruk geven over de vraag in hoeverre de school op niveau zit, en hoe zich dat verhoudt tot andere scholen. De meesten vinden de rapportage met de percentielscores echter overzichtelijker en informatiever. Een rapportage met percentielscores waarbij ook gewerkt wordt met kleuren is dan een mogelijkheid. Verder geven de scholen ook voor de schoolrapportages aan dat zowel een absoluut als een relatief referentiepunt gewenst is.

Voor scholen ligt het ook aan het doel waarvoor de schoolrapportage gebruikt wordt welk type score de voorkeur heeft. Zo is op bestuursniveau teveel detailinformatie niet nodig en geeft de indeling in drie categorieën een goed beeld van hoe de school ervoor staat. Als de directie de schoolrapportage echter onder de aandacht van de secties wil brengen, dan is een verdeling in percentielscores op de verschillende deelaspecten van grotere waarde.

Relatieve of absolute grenswaarden voor categorieën
De meeste scholen geven aan dat zowel een relatieve als een absolute norm interessant is. Bij voorkeur is er sprake van een combinatie: zowel een relatieve lat, zodat de school weet hoe de leerling ervoor staat ten opzichte van het landelijk gemiddelde, als een absolute lat, zodat duidelijk is hoe de leerling het doet ten opzichte van een inhoudelijk vastgestelde norm. Als scholen moeten kiezen geeft de meerderheid de voorkeur aan een absolute lat.

Conclusie

In dit hoofdstuk is op basis van de afnamegegevens van de DTT pilot een continue indicator geïntroduceerd, voor zowel individueel niveau als geaggregeerd op schoolniveau. De precisie van

[30] Dit hangt ook weer af van het doel van de toets. Als de toets bedoeld is als eenmalig tussenmeetmoment richting het eindexamen, dan is een absolute standaard die informatief is over of de leering op koers ligt voor dat examen een logische keuze. Wanneer de toets (of toetsen) deel uit maken van een continue evaluatief proces, dan is herhaalde afname cruciaal en is een relatieve standaard een meer praktische oplossing.

deze score is op individueel niveau relatief laag voor deelaspecten van de toets, maar hoog op het niveau van hoofdaspecten en het overkoepelende vak. Op schoolniveau wordt ook voor deelaspecten een hoge precizie bereikt. Op het niveau van de school worden namelijk de gegevens van vele leerlingen samengenomen, waardoor de meetfout afneemt. Ten opzichte van het samenvoegen op schoolniveau van de categoriale indicator van de DTT is deze continue indicator preciezer, omdat deze ook verschillen binnen categorieën meeneemt. Vergelijkingen laten zien dat de relatieve prestaties van een school duidelijk kunnen verschillen tussen beide indicatoren. Het meenemen van verschillen *binnen* de categorieën onder, op of boven niveau kunnen dus veel uitmaken voor conclusies over het niveau van een school op bepaalde aspecten van de toets. Naast dit psychometrische voordeel van de continue indicator, wordt ook door de bevraagde scholen aangegeven dat de continue indicator, vertaald in een percentielscore, een informatiever beeld geeft over hoe de school ervoor staat op de verschillende onderdelen.

5. Schoolfeedback

In het vorige hoofdstuk is aangetoond hoe er op basis van de afnamegegevens van de DTT pilot een continue score geconstrueerd kan worden op leerlingniveau, die vervolgens geaggregeerd kan worden op schoolniveau. In dit hoofdstuk bekijken we hoe we op basis van deze alternatieve score ook alternatieve schoolrapportages kunnen ontwikkelen. Deze rapportages zijn vervolgens voorgelegd aan de twaalf Limburgse scholen die voor het doel van dit onderzoek zijn bezocht.

5.1 Mogelijkheden voor indicatoren voor schoolfeedback

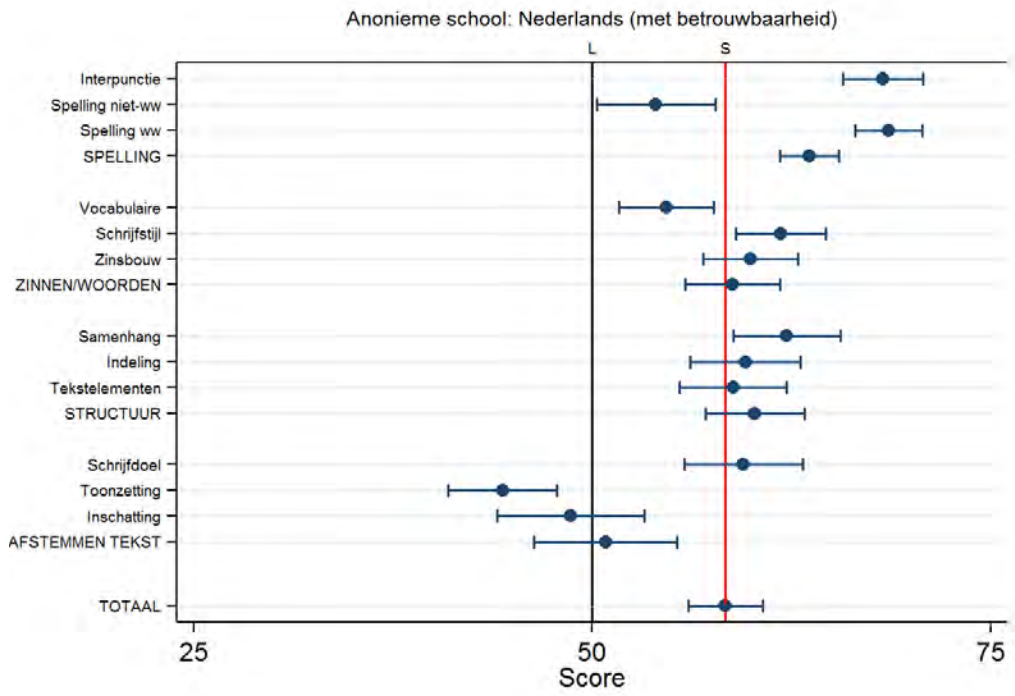
Het doel van de schoolrapportages is om scholen inzicht te geven in hun sterke en zwakke punten. Voor een school zal het zowel interessant zijn om te zien hoe ze gemiddeld scoren als hoe de spreiding van scores binnen de school ligt. Scholen zouden ook geïnteresseerd kunnen zijn in de betrouwbaarheid van de verschillende scores. Bij het tonen van deze informatie is de keuze van het referentiepunt ook een belangrije keuze. Willen scholen zich vooral vergelijken met het landelijke gemiddelde, of zijn ze eerder geïnteresseerd in een vergelijking van de deelaspecten ten opzichte van de prestaties op het overkoepelende onderdeel? We bekijken een aantal mogelijke alternatieven voor het leveren van de schoolrapportages op basis van de continue scores. Hiervoor gebruiken we als voorbeeld de verschillende scores van Nederlands. De voorbeelden zijn van een anonieme VWO-school die in 2017 aan de DTT pilot heeft deelgenomen.

Een andere keuze ligt in de uitkomstmaat. In het vorige hoofdstuk is een gestandaardiseerde continue score geintroduceerd, met een gemiddelde van 0 en een standaard deviatie van 1. Voor scholen zullen verschillen in standaard deviaties waarschijnlijk intuïtief weinig informerend zijn. We gebruiken daarom voor alle schoolrapportages hieronder een percentielscore als uitkomst. Deze score drukt eerst alle individuele scores uit in percentielen van 1 tot 100 (apart berekend per onderwijsniveau) en neemt vervolgens op schoolniveau het gemiddelde van deze percentielscores. Een waarde van bijvoorbeeld 60 betekent dat de gemiddelde leerling op dit onderwijsniveau van deze school landelijk gezien in het 60^{e} percentiel zit. De gemiddelde leerling op de school presteert in dat geval duidelijk beter dan de gemiddelde leerling in Nederland op dit onderwijsniveau. Het landelijk gemiddelde ligt, per definitie, op 50.

5.1.1 Gemiddelde scores en betrouwbaarheid

We tonen allereerst de gemiddelde scores voor de voorbeeldschool, op alle onderdelen van DTT Nederlands. Deze zijn zichtbaar in Figuur 15. Onderaan de figuur is de totale score te zien, de vier hoofdaspecten (Afstemmen van teksten, Tekststructuur, Zinnen/Woorden en Spelling) staan in hoofdletters aangegeven en elk van de drie deelaspecten van de hoofdaspecten staan boven elk hoofdaspect. De figuur trekt verticale lijnen bij het landelijke gemiddelde (bij het 50^{e} percentiel, gelabeld met een ‘L’ aan de bovenkant van de figuur) en bij het gemiddelde van de school voor Nederlands als geheel (gelabeld met een ‘S’ aan de bovenkant van de figuur). De voorbeeldschool hier presteert gemiddeld zeer goed. Voor twee deelaspecten (‘Toonzetting’ en ‘Inschatting’) zit deze school echter onder het landelijk gemiddelde. Een vergelijking met de rode lijn laat zien op welke onderdelen op deze school relatief beter en slechter wordt gepresteerd. Voor veel onderdelen zit deze school net boven het eigen gemiddelde voor Nederlands als geheel, voor een aantal ver erboven (spelling van niet-werkwoorden en interpunctie) en voor een aantal onderdelen redelijk tot ver eronder (Vocabulaire, Spelling van niet-werkwoorden en met name Toonzetting en Inschatting).

Figuur 15: schoolrapportage; gemiddelde en betrouwbaarheidsinterval



De figuur geeft ook de 95%-betrouwbaarheidsintervallen; i.e. de bandbreedte in scores waarvan we met 95% zekerheid kunnen zeggen dat het werkelijke niveau daar tussen ligt. De *schatting* van de schoolscore voor Nederlands als geheel is hier ongeveer 58 (weergegeven door de verticale rode lijn); het *betrouwbaarheidsinterval* geeft aan dat we 95% zeker weten dat de schoolscore binnen percentiel 56 en percentiel 60 valt (weergegeven door de horizontale lijnen). De precisie op schoolniveau is hier dus zeer hoog. Aan alle scholen wordt de figuur eerst alleen getoond met de gemiddeldes, en daarna ook met de betrouwbaarheid. In Sectie 5.3 wordt besproken in hoeverre betrouwbaarheid van de schoolscores voor scholen iets is waar ze behoefte aan hebben of rekening mee houden.

Voor de analyse hier is de betrouwbaarheid in ieder geval een belangrijk kenmerk. Gemiddeld gezien zijn de intervallen smaller voor de totaalscore en de hoofdaspecten, wat verwacht mocht worden gegeven het feit dat er op deze niveaus informatie voor meer vragen meegenomen wordt. De verschillen tussen hoofdaspecten en deelaspecten zijn echter relatief klein, en soms is het interval zelfs breder voor het hoofdaspect (zie bijvoorbeeld ‘Afstemmen tekst’ vergeleken met ‘Toonzetting’). Dit kan voorkomen wanneer scholen sterk verschillend presteren op deelaspecten binnen een hoofdaspect. In dat geval is er automatisch meer onzekerheid rond de nauwkeurigheid van het onderliggende hoofdaspect.

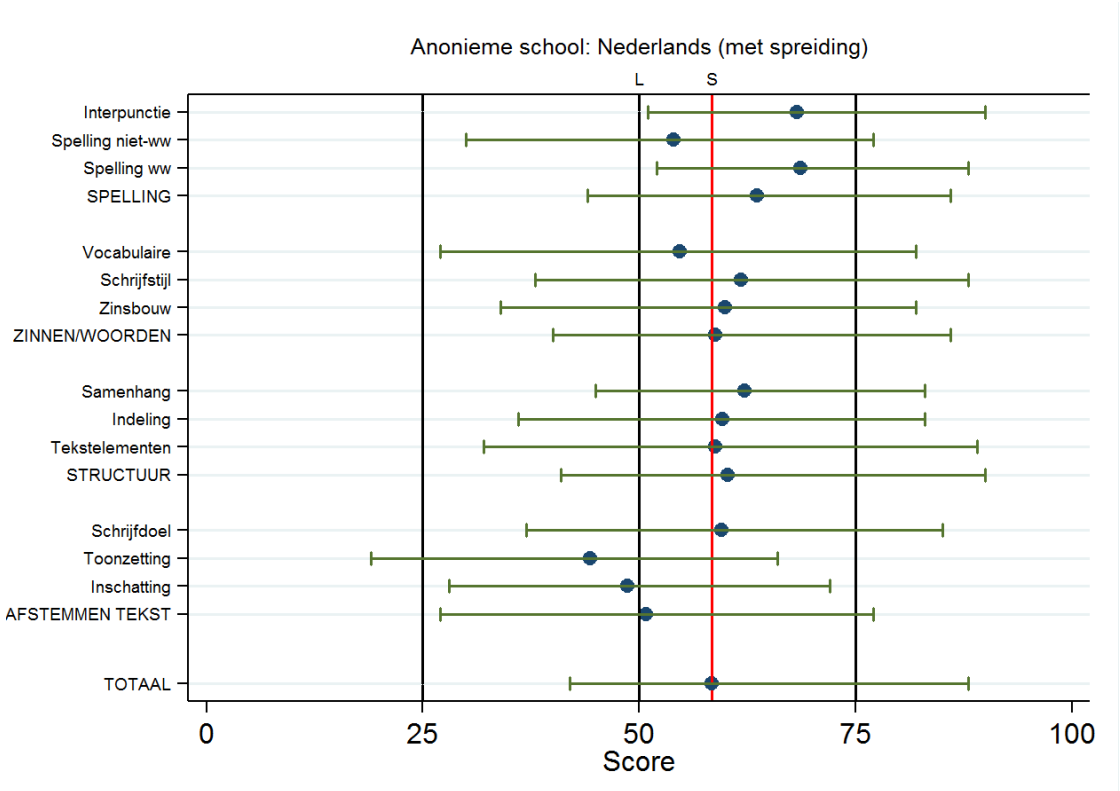
De betrouwbaarheidsintervallen kunnen voor scholen nuttig zijn om te weten hoe zeker het is dat leerlingen op bepaalde onderdelen een verschillend niveau hebben. Zo scoren de leerlingen op deze school hoger op ‘Indeling’ dan op ‘Tekstelementen’, maar gegeven de onnauwkeurigheid van de toets kunnen we niet met zekerheid zeggen of dit een werkelijk verschil vertegenwoordigt, of dat dit een ‘toevallig’ verschil is. We kunnen wel met genoeg zekerheid zeggen dat het niveau van de leerlingen op ‘Vocabulaire’ verschilt van het niveau op ‘Schrijfstijl’, of dat het niveau op de ‘Spelling van niet-werkwoorden’ (sterk) verschilt van dat van ‘Spelling van werkwoorden’. In die gevallen is er

namelijk geen overlap in de intervallen. De figuur laat ook zien hoe zeker het is dat er een afwijking is van het landelijke gemiddelde (door te kijken of er overlap is met de verticale zwarte lijn), of dat er op de deelaspecten een afwijking is ten opzichte van het gemiddelde van de school voor Nederlands als geheel (door te kijken of er overlap is met het betrouwbaarheidsinterval van de totaalscore).

5.1.2 Spreiding van de scores

Een voordeel van de originele schoolrapportages, met de aandelen onder/op/boven niveau (zie Figuur 12), is dat scholen kunnen zien hoe de scores verdeeld zijn en daarmee ook een inschatting kunnen maken van verschillen in prestaties binnen de school. De gemiddelde scores verhullen deze informatie. Figuur 16 geeft een alternatieve schoolrapportage waarin zowel gemiddelde als spreiding is opgenomen. De linkergrens van de horizontale groene lijnen geeft aan waar binnen deze school het prestatieniveau ligt van het 25^{e} percentiel en de rechtergrens geeft aan waar het prestatieniveau ligt van het 75^{e} percentiel. Anders gezegd: stel dat er op deze school 100 leerlingen zitten en dat we deze rangschikken van 1 (laagst presterend) tot 100 (hoogst presterend) op basis van hun toetsprestatie. De linkergrens van de groene lijn geeft dan de score weer van de leerling die op deze ranglijst de 25^{e} plaats inneemt, en de rechtergrens geeft dan de score weer van de leerling die op deze ranglijst de 75^{e} plaats inneemt. De landelijk equivalenten van deze leerlingen zitten per definitie op percentiel 25 en percentiel 75, waar er ook twee verticale lijnen zijn getrokken (er zijn ook wederom verticale lijnen getrokken bij het landelijke gemiddelde op 50 en het gemiddelde van de school voor Nederlands als geheel).

Figuur 16: schoolrapportage; gemiddelde en spreiding



Wanneer de groene lijn voor de school verder naar links loopt dan de verticale lijn bij 25, dan betekent dit dat de 25^{e}-percentiel leerling op deze school zwakker presteert dan de landelijke 25^{e}

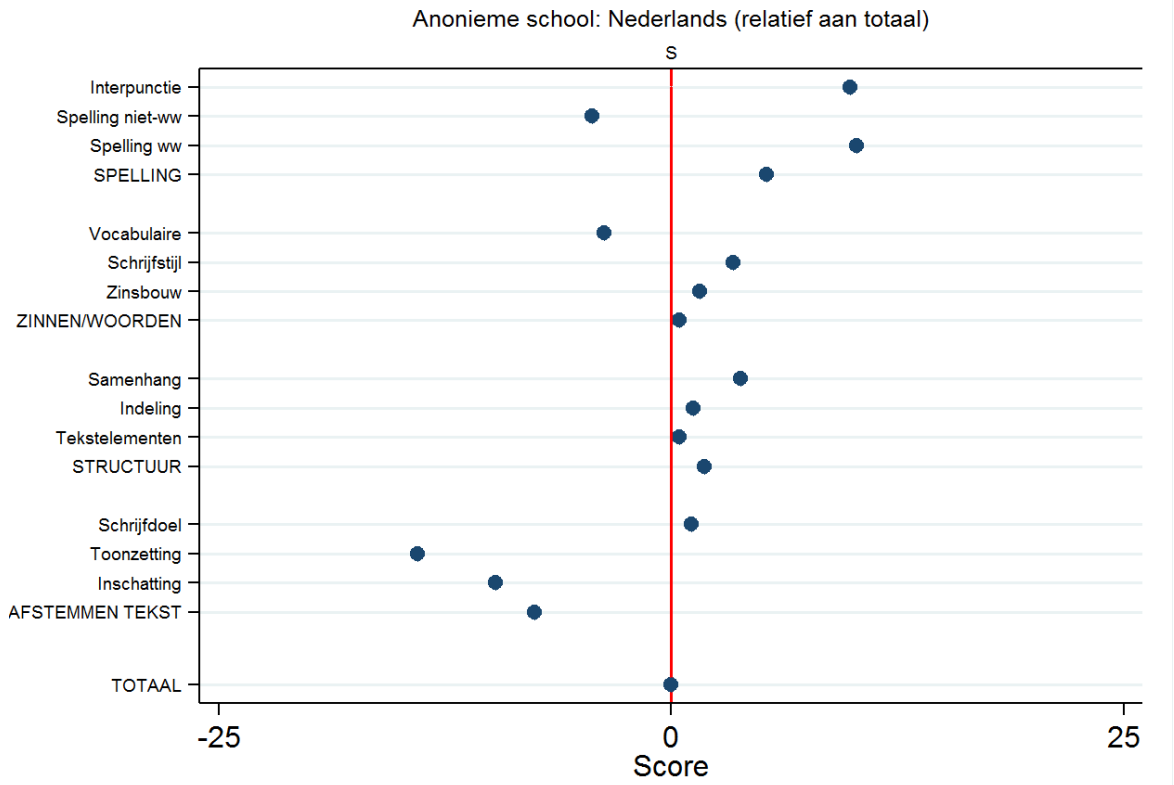
percentiel leerling. Wanneer de horizontale lijn voor de school verder naar rechts loopt dan de verticale lijn bij 75, dan betekent dit dat de 75e-percentiel leerling op deze school het beter doet dan de landelijke 75e percentiel leerling. Ruwweg gezegd betekent het in het eerste geval dat de 'minder presterende leerlingen' op deze school het slechter doen de 'minder presterende leerlingen' vanuit landelijk perspectief, en in het tweede geval dat de 'beter presterende leerlingen' op deze school het beter doen de 'beter presterende leerlingen' vanuit landelijk perspectief. De school in het voorbeeld presteert bijvoorbeeld op het onderdeel 'Interpunctie' zo sterk dat zelfs de relatief zwakkere leerlingen boven het landelijke gemiddelde zitten (de linkergrens blijft immers rechts van het landelijke gemiddelde. Daarnaast blijkt voor het onderdeel 'Inschatting' dat het gemiddelde van de school weliswaar onder het landelijk gemiddelde ligt, maar dat de 'minder presterende' leerlingen op deze school (het 25e percentiel dus) het wel relatief beter doet ten opzichte van het landelijke patroon.

Verder is het ook interessant om te kijken hoe breed de groene lijn is. Hoe breder de lijn, hoe meer variatie er is in de prestaties van leerlingen van deze school. Zo zien we hier bijvoorbeeld dat de leerlingen op 'Samenhang' dichter bij elkaar zitten qua scores vergeleken met het onderdeel 'Tekstelementen'. Op het laatste onderdeel is er dus meer ongelijkheid op deze school, en dus wellicht ook meer reden tot differentiatie. Tot slot zien we voor de totaalscore dat de spreiding naar rechts toe sterker is dan de spreiding naar links toe. Dit betekent dat er relatief meer leerlingen net onder de gemiddelde score zitten en relatief minder leerlingen net boven de gemiddelde score zitten.

5.1.3 Keuze van het referentiepunt

In de hiervoor besproken rapportages is de uitkomstmaat gebaseerd op de gemiddelde ranking van leerlingen ten opzichte van alle leerlingen voor dat onderwijsniveau. Het referentiepunt is dus de gemiddeld presterende leerling op hetzelfde onderwijsniveau (het 50e percentiel). Er kan ook gekozen worden voor andere referentiepunten. Een eerste alternatieve optie is om niet naar het landelijk gemiddelde te kijken, maar om voor de school zelf te kijken hoe de verschillende deelaspecten relateren aan de prestatie voor Nederlands als geheel. Figuur 17 geeft een voorbeeld, voor dezelfde school. De totaalscore staat logischerwijs op nul. Ten opzichte van de eerdere figuren schuiven de scores van deze school dus op naar links. De school kan hiermee dus directer zien in hoeverre het niveau op de deelaspecten van Nederlands voor- of achterblijft ten opzichte van het algemene niveau voor Nederlands.

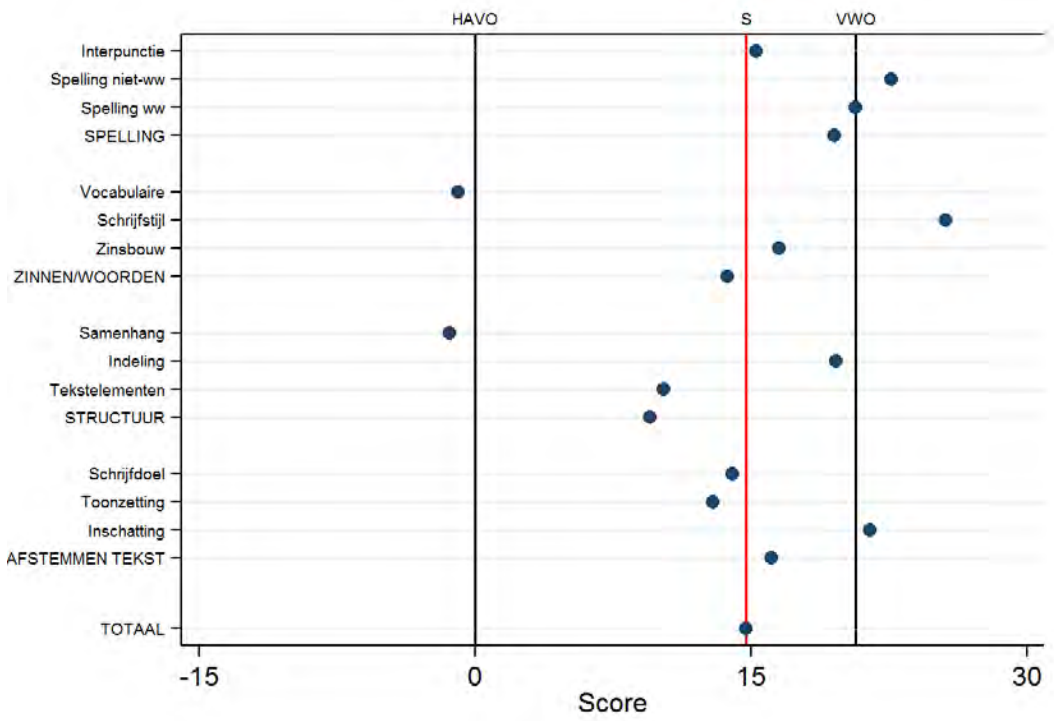
Figuur 17: schoolrapportage relatief aan eigen overkoepelende score



Voor de relatief beter en relatief minder presterende scholen kan het verder ook interessant zijn om een vergelijking te maken met de gemiddelde school op een lager onderwijsniveau of met de gemiddelde school op een hoger onderwijsniveau. Voor een goede havo-school kan bijvoorbeeld aanvullend gekeken worden hoe ze presteren ten opzichte van de gemiddelde vwo-school (of eventueel ook ten opzichte van een laag presterende vwo-school), en voor een slecht presterende vwo-school kan de gemiddelde havo-school een interessante referentie zijn. Wanneer bijvoorbeeld een vwo-school op spelling achter blijkt te liggen op zelfs een typische havo-school dan is dit een extra signaal dat dit een onderdeel is voor deze school om aan te werken. Figuur 18 geeft een voorbeeld voor een relatief zwak presterende vwo-school (dit is een andere school dan in de voorgaande voorbeelden). We vergelijken deze school met de gemiddelde scores voor havo-leerlingen, die hier op 0 zijn gezet.[31] Voor de volledigheid is ook de vergelijking met de gemiddelde vwo-school gegeven. Hoewel deze school dus lager dan gemiddeld scoort vergeleken met andere vwo-scholen, presteren ze voor Nederlands als geheel nog steeds duidelijk beter dan de gemiddelde havo-school. Voor de onderdelen Samenhang en Vocabulaire zitten ze echter ook onder het havo-gemiddelde. Voor de school in kwestie kan dit een meer concrete maat zijn van de mate van achterstand, ten opzichte van een achterstand uitgedrukt in percentielen. Het feit dat de school op bepaalde onderdelen 'zelfs' achterblijft bij het gemiddelde havo-niveau kan daarbij een buitengewoon sterk signaal zijn om het onderwijs op dit onderdeel bij te sturen.

[31] Het vergelijken van havo-leerlingen en vwo-leerlingen is mogelijk, omdat er overlap zit in de gemaakte vragen. In het IRT model kan er daarom een vergelijkbare score geconstrueerd worden.

Figuur 18: schoolrapportage, vwo-school ten opzichte van gemiddelde havo



Wanneer een school de toets vaker afneemt kan ook de prestatie van de school bij de vorige afname gekozen worden als referentiepunt in de schoolrapportage (en deze score voor alle onderdelen dus op nul worden gezet). Hierbij kunnen dan ook de betrouwbaarheidsintervallen worden gegeven, waardoor ook gezien kan worden of de verandering in de prestatie statistisch significant is.

5.2 Mogelijke veranderingen in de opzet van de toets voor de verbetering van deze functie

De DTT is primair opgezet als een toets voor het meten van vaardigheden op leerlingniveau. In dit rapport ligt de nadruk voor een groot deel op het gebruik van de gegevens uit de DTT pilot voor indicatoren op schoolniveau. Het feit dat de toets hier niet primair voor is ontwikkeld kan bepaalde beperkingen impliceren voor het gebruik op schoolniveau. Dit betekent dus ook dat bepaalde aanpassingen aan de toets de schoolfunctie eventueel verder kunnen verbeteren. Een concreet voorbeeld is de selectie van de items. Voor het individuele niveau zouden we idealiter vooral vragen willen stellen die een hoog onderscheidend vermogen hebben rond de grenswaardes. Zeer moeilijke of zeer makkelijke vragen zijn daarbij minder informatief. Voor het schoolniveau zouden we eerder een gelijkmatig spreiding aan vragen willen hebben.

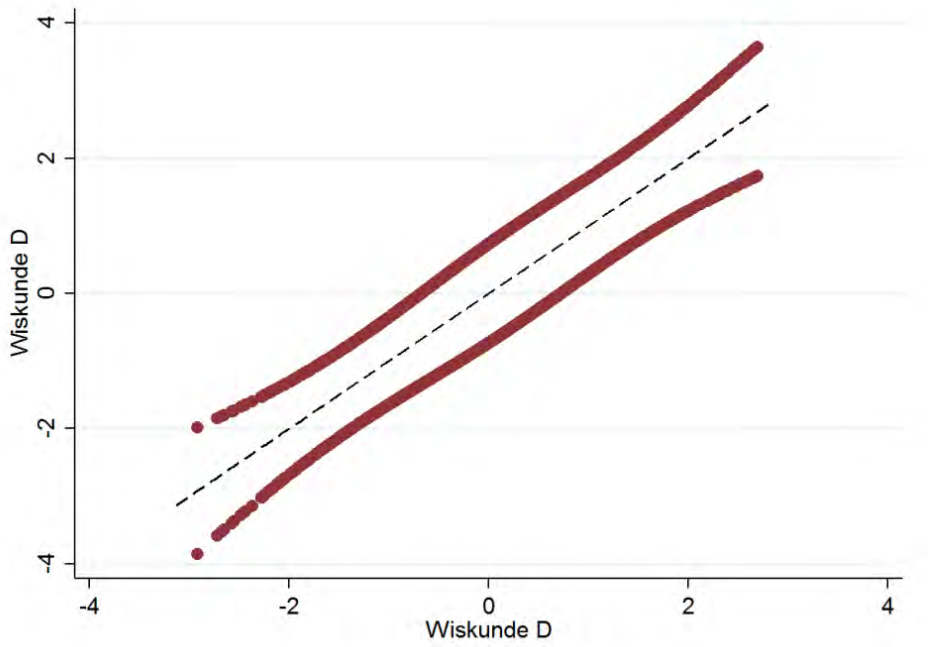
De Figuren 19 en 20 laten een oefening zien waarbij we extra vragen rondom de grenswaarde simuleren.[32] Wanneer we dit doen voor onderdeel D van wiskunde, dan is de impact hiervan zeer beperkt. Dit komt met name omdat het discriminerend vermogen van de vragen niet heel sterk is.

[32] Deze simulatie is uitgevoerd op basis van 1800 observaties. Op basis van het spectrum van geschatte 2PL IRT parameters uit het specificieke domein van de DTT (geschat op basis van de afnamedata van de 2016 afname van de DTT) trekken we willekeurig de waarde voor deze parameters in de simulatie. Voor de extra vragen doen we dit ook wat betreft het onderscheidend vermogen van de vraag, maar zetten we de moeilijkheid vast op één van de twee grenswaardes.

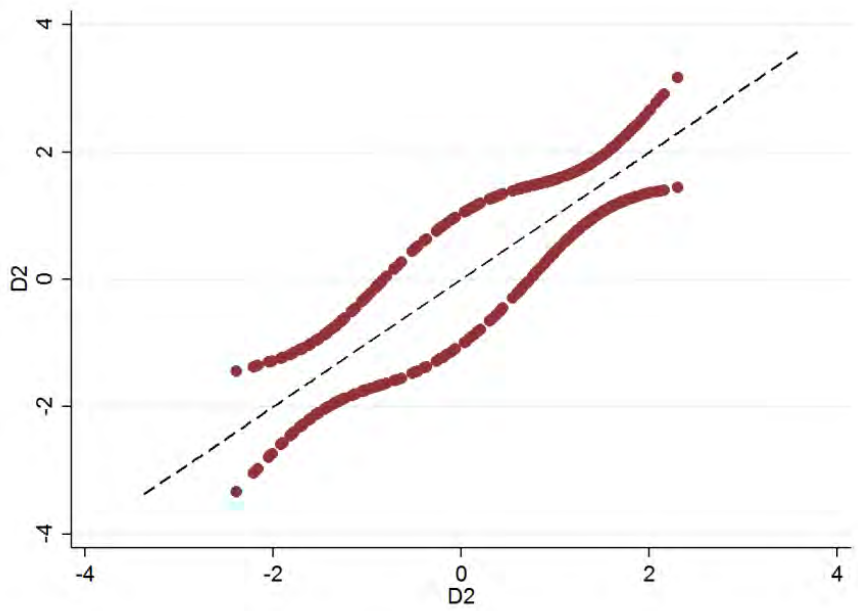
Omdat deze vragen vrij gelijkmatig onderscheidend zijn krijgen we ook bij vragen die we specifiek selecteren om een moeilijkheidsgraad te hebben rond de grenswaarde nog steeds informatie over het hele spectrum.

In het specifieke geval van onderdeel D2 van wiskunde hebben de vragen een zeer hoog onderscheidend vermogen. In dat geval zien we dat het toevoegen van vragen die onderscheidend zijn rond de grenswaarde wel uitmaakt voor de precisie rond die punten. In dat geval kan er dus wel een afweging zijn tussen het individuele niveau en het schoolniveau. Bij de meeste onderdelen van de DTT (waaronder onder andere D1 en D3) is de schets van Figuur 19 echter representatief, en is er dus maar beperkt sprake van een dergelijk spanningsveld.

Figuur 19: simuleer vragen rond grenswaarde (wiskunde D)



Figuur 20: simuleer vragen rond grenswaarde(wiskunde D2)



5.3 De kijk van scholen op de alternatieve schoolrapportages

Betrouwbaarheidsintervallen en spreidingsmaat
Sommige scholen geven aan dat de rapportage met betrouwbaarheidsintervallen wel interessant kan zijn, mits er een goede toelichting bij is. De meeste betrokkenen in de school zullen namelijk moeite hebben met het lezen van de intervallen. De meerderheid geeft echter aan dat deze rapportage statistisch ingewikkeld is en bovendien tot verwarring kan leiden. Als er te veel overlap te zien is in de intervallen bestaat het risico dat de scores niet serieus genomen worden.

Ook ten aanzien van de spreidingsmaat in de rapportage geven de meeste scholen aan dit wel interessant te vinden, maar dat het voor de meeste betrokkenen te ingewikkeld zal zijn. Eén van de scholen gaf aan dat de rapportage met de spreidingsmaat wel nuttig kan zijn in het aangeven van hoe groot de noodzaak is binnen een klas om te differentiëren.

Keuze van referentiepunt
Als scholen moeten kiezen tussen het landelijke gemiddelde of het eigen schoolgemiddelde als referentiepunt, dan geven vrijwel alle scholen de voorkeur aan eerstgenoemde. De rapportage waarbij het eigen schoolgemiddelde op de overkoepelende vaardigheid als referentiepunt wordt genomen heeft voor hen weinig toegevoegde waarde. In de rapportage is hetzelfde scorebeeld op deelaspecten te zien als in de rapportage met het landelijk gemiddelde als referentiepunt. Juist het kunnen afzetten van de eigen scores tegenover het landelijk gemiddelde is relevant voor een duiding van de eigen scores. Bij gebruik van het eigen schoolgemiddelde als referentiepunt is het lastig te interpreteren wat de scores nu eigenlijk betekenen. De relatief sterke punten op deelaspecten kunnen landelijk gezien namelijk nog steeds onder het gemiddelde liggen.

Sommige scholen geven aan dat zij het liefst voor beide referentiepunten een rapportage zouden ontvangen. Op managementniveau is het landelijk gemiddelde als referentiepunt echt nodig. Op

sectieniveau zou de vergelijking met het eigen overkoepelende gemiddelde ook interessant kunnen zijn, omdat de sectie zo kan zien op welke punten het meest gestuurd moet worden.

De verschillende referentiepunten die zijn gepresenteerd in de Figuren 15 en 17 zijn ook toegepast op de leerlingrapportages en voorgelegd aan de scholen. Op leerlingniveau zijn de geluiden uit de scholen wisselender als het gaat over de vraag welk referentiepunt de voorkeur heeft: het landelijk gemiddelde als referentiepunt of het eigen gemiddelde op de overkoepelende vaardigheid. Voor de leerling zelf vinden de scholen de rapportage met het eigen gemiddelde als referentiepunt het meest interessant. Zo heeft elke leerling sterkere en zwakkere punten, ongeacht hoe de leerling ervoor staat ten opzichte van het landelijk gemiddelde. Dit kan motiverend werken. Het brengt echter ook het risico met zich mee dat de leerling denkt dat het al wel goed zit op de relatief sterke punten, terwijl deze landelijk gezien onder niveau zouden kunnen zijn. Daarom vinden scholen het verstandig ook de rapportage met het landelijk gemiddelde als referentiepunt mee te nemen. Vooral voor de docent en sectie is deze informatie van belang. Maar ook in de richting van ouders zouden scholen kiezen voor de leerlingrapportage waarin het landelijk gemiddelde als referentie is genomen. Richting ouders kan deze informatie ook behulpzaam zijn als extra, onafhankelijk gegeven bij het determineren van hun kind.

Tot slot is er nog een optie aan scholen voorgelegd om de gemiddelde score van scholen op een hoger/lager onderwijsniveau als referentiepunt te nemen. De meeste scholen gaven aan weinig behoefte aan deze rapportage te hebben.

Terminologie in rapportages

Tijdens de gesprekken is ook gepeild bij de scholen welke terminologie of manieren van weergave van scores in rapportages hun voorkeur heeft. Percentielscores blijken voor de meeste scholen herkenbaar en goed te lezen. Scholen zijn gewend aan staaf- en circeldiagrammen om resultaten weer te geven, wat een overweging kan zijn om de vormgeving hierop aan te passen. Het werken met standaarddeviaties heeft niet de voorkeur, dit is te moeilijk voor de meeste betrokkenen in een school.

Conclusie

In dit hoofdstuk zijn de continue schoolscores uit Hoofdstuk 4 gebruikt om alternatieve schoolrapportages samen te stellen, op basis van de afnamegegevens van de DTT pilot. We ontwikkelen en presenteren een aantal varianten, die zowel gemiddelde prestaties, betrouwbaarheid van de inschatting en spreiding van de prestaties weergeven. Ook presenteren we alternatieve keuzes voor het referentiepunt in de rapportage. Uit de reacties van de bezochte scholen blijkt dat deze rapportages als informatief worden gezien. Sommige van de alternatieve rapportages wordt door een merendeel als te technisch ervaren, dus het is belangrijk dat deze een geode balans behouden tussen informatieve waarde en overzichtelijkheid.

Het feit dat de DTT is opgezet vanuit een latent klassemodel dat focust op het leerlingniveau kan in potentie spanningen opleveren met het gebruik van een continue indicator op schoolniveau, vooral in de ideale selectie van de items. Een analyse in dit hoofdstuk laat echter zien dat de consequenties hiervan beperkt zijn, gegeven de huidige kwaliteit van de items. Het toevoegen van vragen die specifiek onderscheidend zijn rond de categoriale grenswaardes doet geen sterke afbreuk aan de precisie van de continue indicator rond andere punten.

6 De relatie tussen toetsscores en leerlingkenmerken

Voor PO-scholen en VO-scholen in Limburg worden sinds 2008 uitgebreide gegevens verzameld binnen de Onderwijsmonitor Limburg (OML). De OML verzamelt onder andere om toetsresultaten, achtergrondinformatie, thuissituatie, motivatie en persoonlijkheid van de leerlingen. Voor leerlingen op Limburgse scholen die ook aan de DTT pilot hebben deelgenomen kan er daarom ook een koppeling gemaakt worden met deze gegevens. In dit hoofdstuk relateren we de DTT-prestaties aan prestaties op de Centrale Eindtoets PO, aan achtergrondgegevens zoals opleiding van de ouders en aan indicatoren van persoonlijkheid en motivatie. Op deze manier wordt duidelijk op welke deelaspecten bepaalde groepen leerlingen relatief zwak scoren en hoe de samenstelling van de leerlingpopulatie samenhangt met de leeropbrengsten van een school. Dit kan waardevolle informatie zijn voor het onderwijsbeleid van scholen.

6.1 Toetsscore en eerdere toetsresultaten

Binnen de OML vindt de dataverzameling in het VO plaats in het 3^{e} leerjaar voor zowel vmbo als havo/vwo. Deze wordt afgenomen in het voorjaar van even kalenderjaren. De DTT pilot is afgenomen in 2-vmbo en in 3-havo en 3-vwo. Dit betekent dus dat er een link gemaakt kan worden met vmbo-leerlingen voor DTT resultaten uit 2015 en met havo/vwo-leerlingen voor DTT resultaten uit 2016. In beide gevallen kan er gekoppeld worden aan de 2016-afname van de OML. Voor de 2017 afname van de DTT kan er nog geen koppeling gemaakt worden, aangezien deze leerlingen pas in 2018 in de OML terug te vinden zijn.

In totaal kan er voor 280 leerlingen een koppeling gemaakt worden met DTT Nederlands, voor 195 leerlingen met DTT wiskunde en voor 43 leerlingen met DTT Engels. In totaal zijn er 515 leerlingen in de OML waarvoor er informatie is van ten minste één DTT toets (3 gekoppelde leerlingen maakten zowel de Nederlands als de wiskunde toets).

We vergelijken allereerst de continue scores die geschat zijn op basis van de afnamedata van de DTT met de scores van de Centrale Eindtoets, gemaakt in groep 8. Tabel 3 geeft gestandaardiseerde regressieresultaten, voor DTT wiskunde. In alle gevallen vormt het DTT onderdeel (in de kolommen) de uitkomstvariabele en zijn de deelscores van de Centrale Eindtoets de onafhankelijke variabelen. De onafhankelijke variabelen zijn gezamenlijk binnen hetzelfde model geschat, dus ze schatten de relatie met het DTT onderdeel conditioneel op de scores op de andere onderdelen van de Centrale Eindtoets. We controleren verder voor onderwijsniveau en voor het type toets (pretest 2015, pretest 2016 of adaptieve toets 2016). De tabel toont aan dat de wiskunde toets van de DTT vooral sterk relateert aan het onderdeel meetkunde van de Centrale Eindtoets. Logischerwijs is dit vooral het geval voor de deelscore op Meetkunde voor de DTT, maar ook voor ‘Verbanden en Formules’ is de voorspellende waarde sterk. Het onderdeel breuken en percentages van de Centrale Eindtoets is voor geen enkel onderdeel van de DTT voorspellend (conditioneel op de andere variabelen). Dit geldt ook voor de verschillende deelaspecten van Meetkunde en Verbanden en Formules. Het onderdeel Getallen en Berekeningen van de Centrale Eindtoets is wel voorspellend voor het DTT onderdeel Verbanden en Formules.

Het is verder opvallend dat de taalprestatie op de Centrale Eindtoets (CET) een negatieve relatie heeft met Meetkunde op de DTT (en ook met de totale wiskunde score, als gevolg hiervan). Dit betekent dat, gegeven de prestaties op de andere onderdelen van de Centrale Eindtoets en het

onderwijsniveau op het VO, leerlingen met een betere taalprestatie op de CET slechter scoren op de wiskunde-toets van de DTT. Er is geen relatie tussen studievaardigheden zoals gemeten op de CET en de wiskunde onderdelen van de DTT.

Tabel 3: DTT wiskunde en de Centrale Eindtoets

CET/DTT	Wiskunde totaal	Meetkunde	Verbanden/ Formules
Getallen/Berekeningen	0.121	0.030	0.172*
Breuken/Percentages	-0.023	-0.021	-0.042
Meetkunde	0.262***	0.263**	0.185*
Taal	-0.163*	-0.237**	-0.050
Studievaardigheid	-0.072	0.013	-0.131

Tabel 4 geeft de relatie tussen de Centrale Eindtoets en DTT Nederlands. Voor alle deelaspecten van de DTT blijkt dat de prestaties voor Schrijven op de Centrale Eindtoets het meest voorspellend is van alle taalonderdelen. Dit is niet verrassend aangezien de DTT Nederlands toets specifiek rond schrijfvaardigheden is opgebouwd. Toch is het opvallend dat Schrijven veel meer voorspellend is voor Spelling en Interpunctie dan het onderdeel Spelling van de CET. Dit geldt ook nog als we nog specifieker opsplitsen. Zowel de DTT als de CET meten apart spelling van werkwoorden en spelling van niet-werkwoorden. Ook voor deze deelaspecten is de geschatte relatie zwakker dan tussen DTT-spelling en CET-schrijven. Mogelijk heeft dit te maken met het feit dat spelling een onderdeel van taal is dat regelmatig bijgehouden moet worden en daardoor wellicht meer variabel is door de tijd heen, terwijl schrijven meer een algemeen onderliggend taalniveau meet. Het zou interessant zijn om deze resultaten te bekijken voor een grotere groep leerlingen. Een vergelijkbare bevinding is zichtbaar voor Woordenschat op de CET, dat geen (conditionele) relatie heeft met Woord/Zinsniveau op de DTT. Dit geldt ook voor alle deelaspecten van Woord/Zinsniveau op de DTT. Hierbij geldt wel dat de invulling van het onderdeel in beide toetsen enigszins verschilt. Woordenschat op de CET is vooral gericht op woordbetekenissen terwijl het DTT-onderdeel vooral gaat om passend woordgebruik afhankelijk van context en om juist gebruik van zinsconstructies.

De resultaten tonen verder significante relaties aan tussen Spelling en Tekststructuur, en tussen Begrijpend Lezen en Afstemmen op Doel en Publiek. Studievaardigheid op de CET blijkt een sterke relatie te hebben met Tekststructuur en Woord/Zinsniveau. In beide gevallen is deze relatie zelfs sterker dan voor het onderdeel Schrijven van de CET. Tot slot is er een negatieve relatie tussen rekenen op de CET en de verschillende DTT Nederlands onderdelen, maar deze geschatte coëfficiënten zijn relatief laag en alleen (marginaal) significant voor Tekststructuur op de DTT.

Tabel 4: DTT Nederlands en de Centrale Eindtoets

CET/DTT	Nederlands totaal	Afstemmen doel/publiek	Tekst structuur	Woord/ zinsniveau	Spelling/ interpunctie
Schrijven	0.290***	0.208**	0.175**	0.124*	0.317***
Spelling	0.122**	0.073	0.111*	0.074	0.105*
Woordenschat	-0.012	-0.058	0.073	0.0043	-0.049
Begrijpend lezen	-0.030	0.167*	0.056	0.039	-0.060
Rekenen	-0.0025	-0.083	-0.153*	-0.118	0.069
Studievaardigheid	0.139*	0.088	0.193**	0.212**	0.086

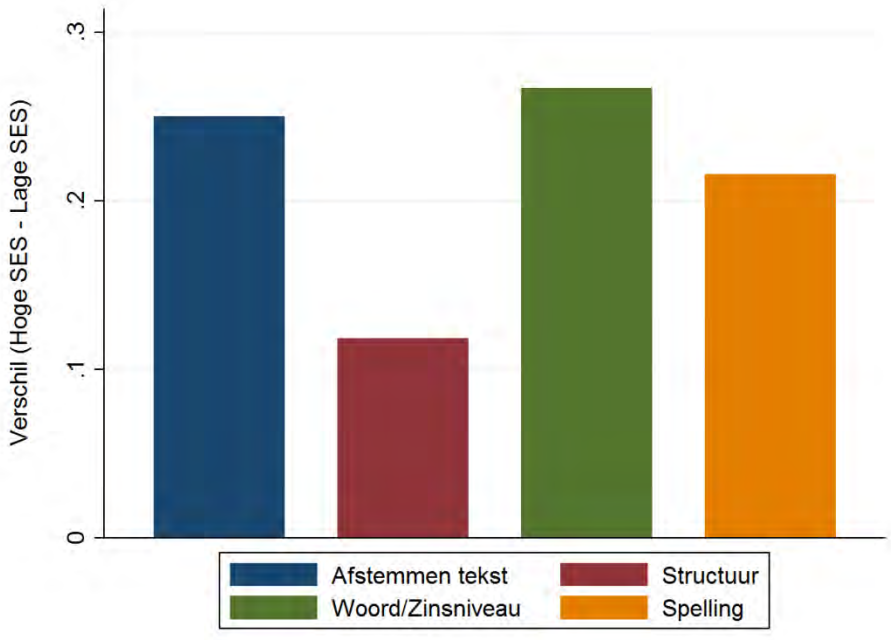
Voor de Engels toets van DTT zijn er te weinig observaties om eenzelfde uitsplitsing te maken. Uit de kleine steekproef blijkt wel dat er een statistisch significante relatie is met zowel Studievaardigheid (positief) en rekenen (negatief).

De prestaties op de DTT correleren logischerwijs ook sterk met de prestaties op de toetsen die in 3VO binnen de OML worden afgenomen. Deze relatie is sterker voor de taaltoetsen dan voor de wiskundetoetsen.

6.2 Toetsscore en achtergrond

Voor de gekoppelde gegevens is het ook mogelijk om de prestaties op de DTT pilot te linken aan achtergrondgegevens van de leerlingen. In de OML wordt onder andere het opleidingsniveau van de ouders gemeten. Uit gegevens van de Centrale Eindtoets weten we bijvoorbeeld dat leerlingen met laag opgeleide ouders vooral slecht scoren op begrijpend lezen en relatief beter presteren op spelling.[33] We bekijken deze relaties hier voor de DTT pilot. In Figuur 22 zien we voor de onderdelen van de DTT Nederlands toets het verschil in prestaties (gecontroleerd voor onderwijsniveau) tussen leerlingen met laag opgeleide ouders en leerlingen met hoog opgeleide ouders (hbo of wo). In alle gevallen scoren de leerlingen van hoog opgeleide ouders beter, maar dit verschil is ruim twee keer zo sterk voor Woord- en Zinsniveau vergeleken met Tekststructuur. Binnen het onderdeel Woord- en Zinsniveau zijn de verschillen vooral sterk rondom het gebruiken van de juiste zinsconstructies en relatief zwakker voor Passend en Gevarieerd Woordgebruik. Binnen Spelling zijn de verschillen relatief kleiner voor de spelling van niet-werkwoorden. Voor de onderdelen van DTT wiskunde zijn er geen sterke verschillen naar sociaaleconomische status (SES) van de leerlingen.

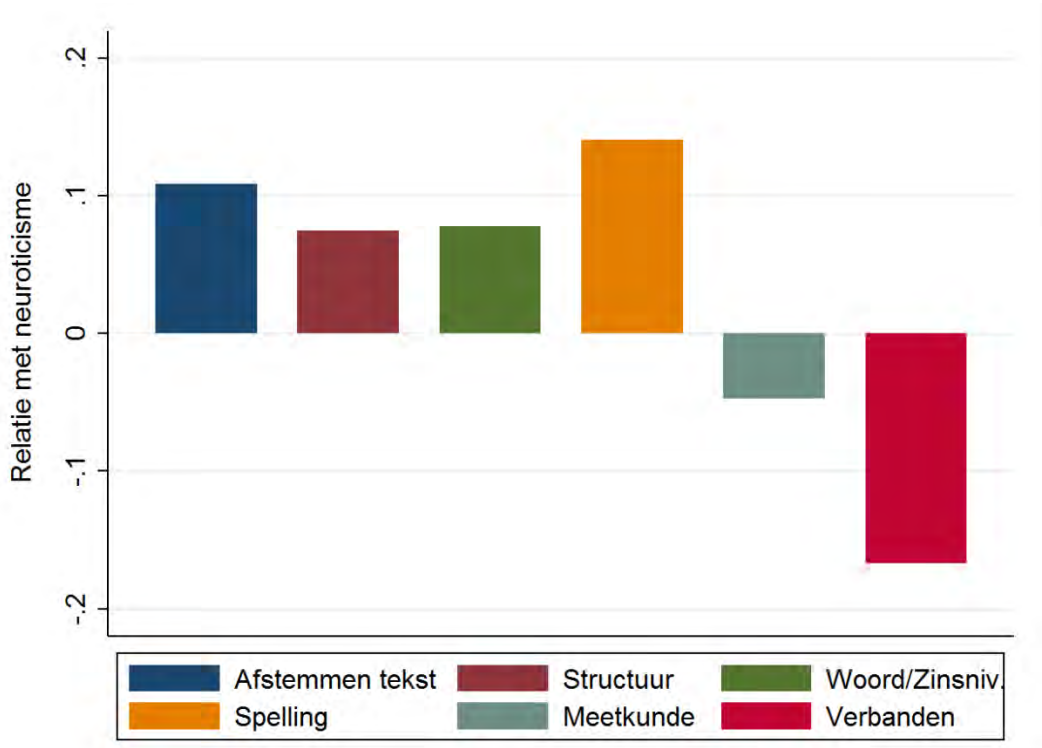
Figuur 21: DTT en sociaaleconomische status (opleidingsniveau ouders)



[33] Zie http://www.educatieveagendalimburg.nl/bijdragen/toetsen/het-belang-van-diagnostisch-toetsen-voor-de-achterstandsproblematiek

De OML bevat verder uitgebreide gegevens over de sociaal-emotionele ontwikkeling van leerlingen, zoals bijvoorbeeld persoonlijkheidskenmerken en schoolmotivatie. Uit eerder onderzoek weten we dat toetsprestaties niet alleen relateren aan cognitieve vaardigheden, maar ook aan non-cognitieve vaardigheden. Voor de gekoppelde leerlingen vinden we ook significante relaties met persoonlijkheid. Hieronder geven we als voorbeeld de relatie met neuroticisme; één van de 'Big Five' persoonlijkheidskenmerken. De figuur geeft de gestandaardiseerde coëfficiënten weer voor de relatie tussen neuroticisme en de onderdelen van DTT Nederlands en DTT wiskunde, gecontroleerd voor onderwijsniveau. We zien dat meer neurotische leerlingen beter presteren op taal, en dan vooral op het gebied van spelling, en slechter presteren op wiskunde, en dan vooral op het gebied van Verbanden en Formules. De andere persoonlijkheidskenmerken zijn relatief minder voorspellend, maar we identificeren onder andere een positieve relatie tussen Zorgvuldigheid en Tekststructuur en tussen Doorzettingsvermogen en Verbanden/Formules.

Figuur 22: De DTT en persoonlijkheidskenmerken



Persoonlijkheid blijkt ook gerelateerd aan het *verschil* tussen de prestaties op de Cito Eindtoets en de prestaties op de DTT. Meer neurotische leerlingen presteren beter op de DTT, wat waarschijnlijk te maken zal hebben met het feit dat de DTT een low stakes toets is en de CET een high stakes toets. Om waarschijnlijk dezelfde reden vinden we een omgekeerde relatie met prestatiegerichtheid. Leerlingen die meer prestatiegericht zijn presteren dus relatief beter op de high stakes CET. Ook blijkt dat leerlingen met een betere schoolmotivatie en een betere schoolhouding in het 3^{e} jaar van het VO (3VO) relatief beter presteren op de DTT ten opzichte van de Centrale Eindtoets. Deze relatie vinden we echter niet wanneer we kijken naar de schoolmotivatie gemeten in groep 8. Het zijn dus de leerlingen die tussen groep 8 en 3VO meer gemotiveerd zijn geraakt voor school die relatief beter presteren op de DTT toets (en vice versa).

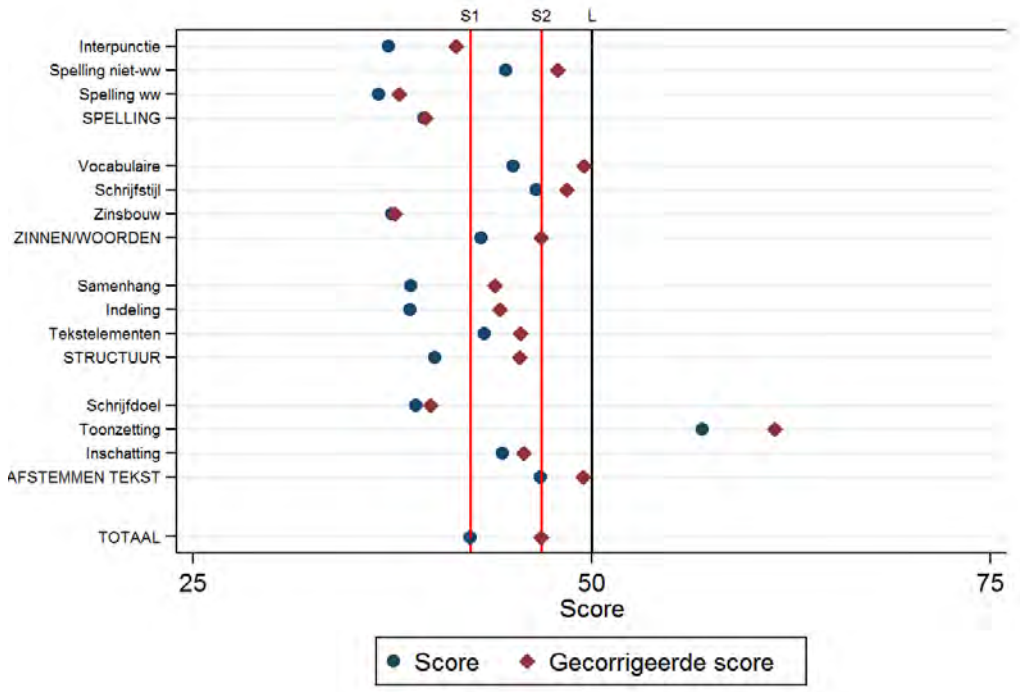
De relatie tussen de prestaties op de DTT pilot en IQ is van een vergelijkbare grootte als de relatie tussen de Centrale Eindtoets en IQ. Beide toetsen relateren dus in vergelijkbare mate met 'pure' cognitie.

6.3 Corrigeren in de schoolrapportages op basis van de gekoppelde data

De schoolrapportages gepresenteerd in Hoofdstuk 5 kunnen ook worden uitgebreid op basis van deze achtergrondinformatie. Een VO-school is voor een deel afhankelijk van het niveau waarop de leerlingen presteren op het moment dat ze het VO binnenstromen. Zo kan het voorkomen dat een school in de DTT bijvoorbeeld relatief slecht scoort op spelling vergeleken met de gemiddelde school, maar dat de leerlingen ook op de Centrale Eindtoets PO al achterlagen op dit onderdeel. Hoewel dit nog steeds redenen geeft om meer te werken aan spelling op basis van de DTT resultaten, geeft het aan dat de oorsprong van de mindere prestaties op 3VO niet ligt in het spellingsonderwijs in de onderbouw van de VO-school. Aan de andere kant kan juist blijken dat op sommige onderdelen in de DTT gemiddeld wordt gepresteerd, maar dat leerlingen hun voorsprong die ze nog op de Centrale Eindtoets hadden verloren zijn. In dat geval is er ondanks de gemiddelde score nog steeds een reden voor deze school om het onderwijzen van dit onderdeel goed onder de loep te nemen. Eenzelfde correctie kan ook toegepast worden voor SES. Gegeven de bestaande link tussen, bijvoorbeeld, SES en begrijpend lezen, zou een lagere score op onderdelen die direct aansluiten op begrijpend lezen voor een bepaalde school bijvoorbeeld deels verklaard kunnen worden door een relatief hoog aandeel leerlingen van lage SES.

Figuur 19 geeft een voorbeeld van hoe een dergelijke correctie er uit zou kunnen zien in de rapportage (dit is een gesimuleerd voorbeeld). De originele scores zijn hier in blauw weergeven (de gemiddelde overkoepelende score is aangegeven met S1) en de gecorrigeerde scores zijn zichtbaar in rood (de gemiddelde overkoepelende score is aangegeven met S2). Deze school presteert dus relatief slecht, maar dit kan voor een deel (maar niet geheel) verklaard worden door relatief minder gunstige achtergrondindicatoren, zoals bijvoorbeeld veel leerlingen met laag opgeleide ouders of leerlingen die op de Centrale Eindtoets gemiddeld ook al lager scoorden op taalonderdelen. Zoals in het voorbeeld te zien is, maakt dit voor sommige deelaspecten meer uit dan voor andere.

Figuur 23: schoolrapportage, met correctie voor achtergrondvariabelen



6.4 De kijk van scholen op de koppeling van de DTT aan aanvullende leerlinginformatie

Het merendeel van de bevraagde scholen vinden het interessant om rapportages op schoolniveau te ontvangen waarin de uitkomst van een diagnostisch instrument zoals de DTT gerelateerd wordt aan leerlingkenmerken en achtergronden. Soms is er binnen de school wel een vermoeden waar een achterstand binnen een bepaald vak vandaan komt, maar is er onvoldoende bewijs in handen om interventies erop te zetten. Door verschillende verbanden in beeld te hebben, ook in relatie tot andere scholen, kunnen vermoedens worden getest en andere verbanden zichtbaar worden. Met die informatie in handen kan de directie discussies op gang brengen in de sectie en kunnen beleidsinterventies ontwikkeld worden. Daarbij is het nuttig als er trends zichtbaar gemaakt kunnen worden over de jaren heen. Eén rapportage op één moment kan een vertekend beeld geven.

Een ander punt dat naar voren werd gebracht binnen dit thema is de meerwaarde van de koppeling van informatie uit de DTT pilot aan vroegere of toekomstige schoolprestaties van de betreffende leerlingen. Bijvoorbeeld: in hoeverre zijn de DTT-resultaten een voorspeller voor eindexamenresultaten? Een belangrijk punt bij scholen op dit moment is afstroom. In hoeverre halen leerlingen met bijvoorbeeld een havo vo-advies ook uiteindelijk een havo diploma? De DTT kan hiervoor een nuttig tussenmeetmoment zijn, waardoor er eventueel kan bijgestuurd worden hierin en daarmee ook de afstroom vermindert.

Andere scholen geven echter aan niet direct behoefte te hebben aan méér rapportages op schoolniveau over toetsscores in relatie tot leerlingkenmerken en achtergrondgegevens. Er ontstaat het risico op een te veel aan informatie, waardoor directies maar ook docenten als het ware kunnen verdrinken in de data.

Corrigeren voor achtergrond

Tijdens de gesprekken is aan scholen het hierboven besproken voorbeeld voorgelegd van een manier waarop de prestaties in de DTT op basis van de continue score gecorrigeerd kunnen worden voor populatieverschillen, in dit geval de SES van leerlingen. De reactie van scholen op dit expliciete voorbeeld is wisselend. Sommige scholen geven aan het relevant te vinden om te corrigeren voor SES, zeker in de richting van het bestuur. Het aannamebeleid ten aanzien van LWOO-leerlingen zal dan bijvoorbeeld zichtbaar zijn in de resultaten. Een correctie geeft het meest eerlijke beeld van hoe de school ervoor staat. Andere scholen geven aan dat dit interessante, maar geen noodzakelijke informatie is. Zij geven aan al te weten hoe de leerlingpopulatie samengesteld is en zijn bekend met het effect hiervan op de toetsscores. Bovendien vragen scholen zich af wat ze kunnen doen met de bevinding, aangezien SES een gegeven is.

Hoewel scholen bekend zullen zijn met het feit dat SES negatief relateert aan schoolprestaties, zal er geen precies beeld zijn over de exacte grootte van deze effecten. Bovendien blijkt ook in dit hoofdstuk dat deze SES-effecten sterk kunnen verschillen naar onderdeel. Om in te schatten in hoeverre SES een verklaring is voor bepaalde tekortkomingen op klas- of schoolniveau is dus zeer precieze informatie nodig, die de scholen niet zullen bezitten.

6.5 Prestaties en tijd besteed per vraag

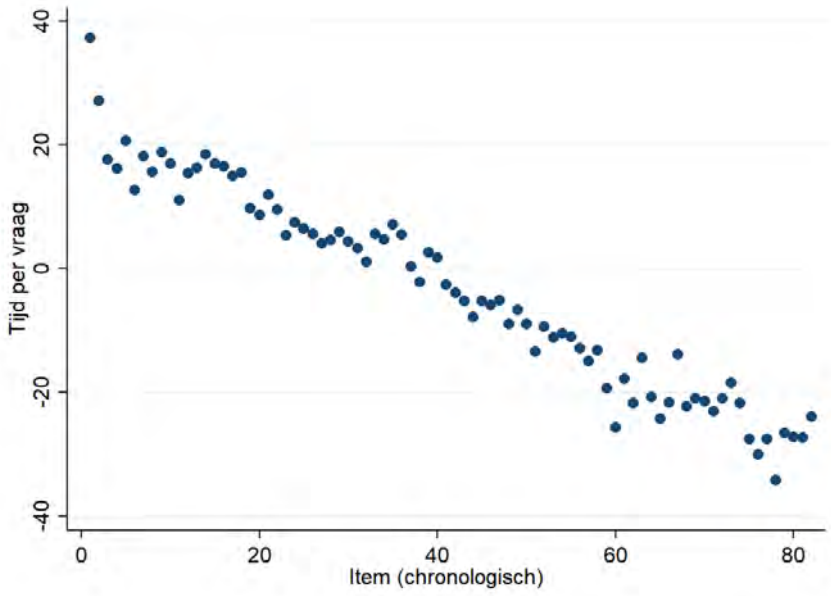
Een voordeel van een digitale toets zoals de DTT is dat er veel gegevens kunnen worden opgeslagen tijdens het maken van de toets. Dit laat toe om meer te analyseren van hoe een leerling een toets maakt dan alleen de toetsscore. Onderzoek heeft aangetoond dat bepaalde informatie die uit toetsen te halen is, zoals bijvoorbeeld de afname in prestaties gedurende de toets, sterk verbonden is met bepaalde persoonlijkheidskenmerken en ook voorspellend is voor latere uitkomsten.

Een typisch voorbeeld van informatie over hoe leerlingen toetsen maken is de tijd die ze besteden per vraag. Deze informatie wordt binnen de DTT pilot precies gemeten, voor elke vraag en elke leerling. We bekijken deze informatie om na te gaan hoe deze bestede tijd zich ontwikkelt gedurende de toets, hoe dit relateert aan prestaties en hoe dit relateert aan achtergrondinformatie van de leerlingen.

Figuur 24 geeft aan hoe de toetstijd zich ontwikkelt voor de Nederlands toets. We gebruiken hierbij de gegevens voor de Pre-test uit 2015. Aangezien deze toets geen adativiteit kent, kunnen we hier een zuiverder beeld krijgen van de tijd per vraag, aangezien de moeilijkheid (en daarmee samenhangend de duur) van de vraag niet afhangt van het niveau van de leerling, binnen zijn onderwijsniveau. De figuur geeft de tijd per vraag afhankelijk van wanneer deze gesteld is. Deze waardes zijn gestandaardiseerd zodat ze de afwijking van het gemiddelde geven voor de specifiek vraag. Een bepaalde vraag kan bijvoorbeeld bij de ene groep leerlingen als 10^{e} zijn gesteld en bij een andere groep leerlingen als 30^{e}. Er wordt dan eerst gekeken naar hoe lang er gemiddeld over die vraag wordt gedaan en dan wordt gekeken hoe veel langer of korter de leerlingen die de vraag als 10^{e} (of als 30^{e}) kregen erover gedaan hebben. Anders gezegd: de waarde bij bijvoorbeeld 10 geeft aan hoe veel langer or korter leerlingen over een vraag doen wanneer deze als 10^{e} wordt gesteld, vergeleken met elk ander punt waarop de vraag gesteld is. Het (gewogen) gemiddelde van alle waardes in de figuur komt dus altijd op 0 uit. In Figuur 24 zien we dat de (gecorrigeerde) tijd per

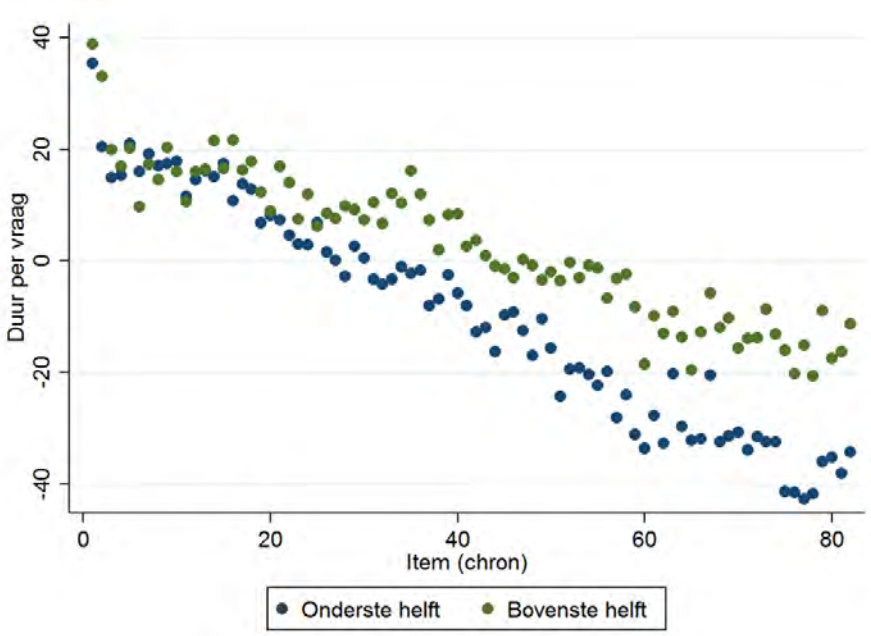
vraag voor Nederlands gestaag afneemt. Bij wiskunde zien we ook een afname, maar daar blijft de tijd langer vlak, en is er een sterke daling naar het einde van de toets toe. Voor Engels is er alleen een daling op het absolute einde. Waarschijnlijk heeft het laatste te maken met het feit dat de Engels toets veel korte (en ook relatief meer gemakkelijke) vragen heeft vergeleken met Nederlands en Wiskunde. Ter vergelijking, de gemiddelde tijd per vraag is ongeveer 79 seconde voor Nederlands, 88 seconde voor wiskunde en 29 seconde voor Engels. Hoewel de Engels toets ook meer vragen heeft, is de gemiddelde totale tijd die de leerlingen aan de toets besteden ook duidelijk lager dan voor Nederlands en wiskunde. Vanuit dat perspectief is het niet verrassend dat de afname in concentratie later komt of zelfs geheel uitblijft.

Figuur 24: bestede tijd per vraag (Nederlands)



We bekijken vervolgens hoe de ontwikkeling in de toetstijd zich verhoudt tot de prestaties op de toets. Hiervoor splitsen we de groep leerlingen in twee delen, op basis van de totaalscore voor het overkoepelende onderdeel. Figuur 25 geeft de ontwikkeling van de bestede tijd gedurende de toets voor beide groepen apart. We zien dat in het begin van de toets de goed en slecht presterende leerlingen nog gelijk liggen in de bestede tijd per vraag. De tijden divergeren echter vrij snel, en tegen het einde van de toets is er een verschil van ongeveer 20 seconde. Voor wiskunde is dit patroon relatief minder duidelijk, maar er is ook daar een positieve correlatie tussen de bestede tijd en de eindscore. Voor Engels is dit juist andersom; beter presterende leerlingen gebruiken juist minder tijd. Dit heeft zeer waarschijnlijk wederom te maken met de relatief korte en makkelijke vragen voor Engels. Aangezien deze vragen een minder zware inspanning vergen, zullen aspecten als de snelheid van denken automatisch meer dominant zijn.

Figuur 25: bestede tijd per vraag en prestatie (Nederlands)



De bovenstaande figuren geven aan dat er, voor de vakken Nederlands en wiskunde, een positieve relatie is tussen de *gemiddelde* tijd besteed per vraag en de prestatie op de toets, en ook dat er een positieve relatie is tussen de *relatieve* afname in bestede tijd per vraag gedurende de toets (i.e. de hellingshoek van de relatie in de figuren hierboven). We zien verder ook dat er een positieve correlatie is tussen de *variatie* in tijd besteed per vraag en de score op de toets. Dit lijkt te suggeren dat de beter presterende leerlingen beter in staat zijn om extra tijd te gebruiken wanneer dat nodig is voor een bepaalde vraag (terwijl ze voor makkelijke vragen juist sneller zijn; zie de bevindingen voor Engels). Bij minder presterende leerlingen lijkt er een bepaald soort 'maximum' aan bestede tijd per vraag te zijn waar ze niet of nauwelijks overheen gaan.[34]

Er zijn ook verschillen in de bestede tijd naar achtergrond van de leerlingen. Meisjes besteden meer tijd per vraag dan jongens (dit is gecontroleerd voor onderwijsniveau). Dit geldt zowel wanneer meisjes beter presteren (Nederlands) en wanneer meisjes minder presteren (wiskunde) dan jongens. In beide gevallen gaat het om ongeveer 5 seconde tijd per vraag meer voor meisjes. Voor Engels is het verschil kleiner; ruim 1 seconde (de prestaties voor jongens en meisjes liggen voor Engels vrijwel gelijk). Er zijn geen consistente verschillen naar onderwijsniveau. We kunnen geen vergelijking maken tussen vmbo aan de ene kant en havo/vwo aan de andere kant, aangezien er geen overlap zit in de vragen, maar binnen het vmbo en tussen havo en vwo is deze overlap wel aanwezig. De verschillen in bestede tijd voor deze overeenkomstige vragen is echter klein, en ook niet consistent over de verschillende vakken. Verder zijn er ook geen verschillen in de relaties tussen prestatie en bestede tijd naar deeldomein.

Conclusie

[34] De bevinding lijkt op het eerste gezicht in contrast met het feit dat het patroon voor de betere leerlingen in figuur 25 juist vlakker is. Echter, figuur 25 geeft tijden die *gecorrigeerd* zijn voor de gemiddelde tijd die leerlingen bij die vraag nodig hebben.

In dit hoofdstuk hebben we een koppeling gemaakt tussen de continue scores die we construeren op basis van de afnamegegevens van de DTT, en de gegevens die aanwezig zijn voor Limburgse leerlingen in de OnderwijsMonitor Limburg (OML). Deze koppeling is voor ruim 500 leerlingen gemaakt, die zowel aan de DTT pilot als de OML deelnemen. Uit de analyses blijkt dat de prestaties op de DTT vooral sterk relateren aan schrijvaardigheid op de Centrale Eindtoets PO. Specifieke deelaspecten van de DTT relateren niet altijd sterk met zeer vergelijkbare deelscores uit de Centrale Eindtoets. Dit suggereert dat zulke vaardigheden veranderlijk zijn door de tijd, en dat een meetmoment halverwege het VO zeer informatief is, zowel op leerlingniveau als schoolniveau. De resultaten laten ook zien dat er verschil is tussen de onderdelen van de DTT in de mate van samenhang met sociaaleconomische status van de leerling. Met name binnen het aspect Woord- en Zinsniveau zijn de verschillen tussen leerlingen van hoge SES en leerlingen van lage SES hoog.

De koppeling van toetsgegevens met achtergrondgegevens van leerlingen kan zeer nuttig zijn om een beter perspectief te krijgen op bepaalde resultaten. Waren slechte scores op bijvoorbeeld spelling al zichtbaar op de Centrale Eindtoets of is deze achterstand pas op het VO ontstaan? Voor zowel leerling als school is dit zeer waardevolle informatie, ook richting het toekomstige leerproces. In de eerder besproken schoolrapportages kan ook op basis van deze gegevens een score worden weergegeven die ‘gecorrigeerd’ is voor SES of voor prestaties op de Centrale Eindtoets PO. Op verschillende van de bezochte scholen wordt dit als een nuttige toevoeging gezien, al moet er tegelijkertijd ook voor gewaakt worden dat dergelijke rapportages te ingewikkeld worden.

7 Het gebruik van de DTT voor gedifferentieerd leren: een analyse van de mogelijkheden

Andrea Oudkerk Pool, en Diana Dolmans
School of Health Professions Education (SHE), Faculty of Health, Medicine and Life Sciences, Maastricht University.

7.1 Inleiding

Steeds meer scholen hebben aandacht voor gedifferentieerd leren. Hierbij wordt het onderwijs aangepast aan de verschillen tussen leerlingen. Om te kunnen differentiëren is het van belang de verschillen tussen leerlingen goed in kaart te brengen en de voortgang van de leerlingen te monitoren. In dit hoofdstuk wordt ingegaan op de mogelijkheden die de DTT zou kunnen bieden voor gedifferentieerd leren.

Omgaan met verschillen tussen leerlingen is een belangrijke uitdaging in het onderwijs. Differentiëren is een aanpak waarbij verschillen tussen leerlingen in kaart worden gebracht en docenten de instructie en begeleiding aanpassen om aan de leerbehoeften van leerlingen tegemoet te komen. Om effectief te kunnen differentiëren zijn verschillende informatiebronnen nodig die het mogelijk maken om de ontwikkeling van de leerlingen te monitoren. Toetsen vormen een van deze bronnen van informatie. Veel scholen maken gebruik van een leerlingvolgsysteem met uniforme toetsen om de ontwikkeling en prestaties van leerlingen in kaart te brengen. Hiervoor wordt gebruik gemaakt van een verscheidenheid aan gegevens zoals scores op methode-onafhankelijke en methode-afhankelijke toetsen, rapportcijfers, gegevens over de sociaal-emotionele ontwikkeling van leerlingen en begeleidingsgegevens. Door analyse van deze gegevens ontstaat een beeld van de ontwikkeling van de leerling over tijd. Behalve analyse op leerling niveau is het in toenemende mate mogelijk om ook op het niveau van de klas en de school de ontwikkeling van leerlingen te analyseren.

De DTT is een diagnostische methode-onafhankelijke toets die gedetailleerde informatie geeft over de prestaties van de leerlingen voor de vakken Nederlands, Engels en Wiskunde. De toets levert informatie op zowel leerling-, groeps-, als schoolniveau.

Het doel van dit hoofdstuk is om inzicht te krijgen in de vraag: Op welke manier kan de informatie uit de DTT data bijdragen aan differentiatie binnen het onderwijs?

Allereerst wordt een overzicht gegeven van de belangrijkste literatuur rondom differentiatie, data-based decision making en opbrengstgericht werken en hoe de verschillende concepten relateren aan het gebruik van de DTT. Op basis hiervan worden aanbevelingen gedaan over hoe de DTT geïmplementeerd zou kunnen worden op een manier dat deze de ontwikkeling van gedifferentiëerd leren optimaal kan stimuleren.

7.2 Literatuuroverzicht

Dit literatuuroverzicht biedt een samenvatting van de belangrijkste inzichten over hoe differentiëren, formatief toetsen en formatieve feedback succesvol geïmplementeerd kunnen worden. Voor elk van deze concepten wordt omschreven in hoeverre de DTT voldoet aan de criteria voor succesvolle

implementatie. Ook wordt aangegeven hoe data-based decision making kan helpen bij het implementeren van de DTT. Tot slot wordt geschetst hoe de DTT past binnen de cultuur van opbrengstgericht werken.

7.2.1 Differentiëren

Differentiëren is een aanpak waarbij de instructie wordt gevarieerd voor en aangepast aan de verschillen tussen leerlingen en de leermogelijkheden van de leerlingen. Deze verschillen worden in kaart gebracht met behulp van systematische procedures voor het volgen van de prestaties en ontwikkeling van leerlingen (Roy et al., 2013).

Differentiëren bevat vier belangrijke componenten (Tomlinson, Brimijoin, & Narvaez, 2008):

- *Doelgericht werken:* Er worden ambitieuze prestatiedoelen nagestreefd die enkel bij hoge uitzondering naar beneden worden bijgesteld. Hoog presterende leerlingen krijgen extra uitdagende doelen.
- *Het in kaart brengen van verschillen:* Dit kan door middel van het vooraf analyseren van de vorderingen van leerlingen (a priori) of door het bepalen van de aanwezige kennis tijdens een les (ad-hoc).
- *Het monitoren van voortgang:* Dit kan door middel van formatieve evaluatie of het stellen van vragen die aansluiten bij het niveau van de leerling.
- *Het aanpassen van de instructie:* Er bestaan verschillende strategieën gericht op het tegemoetkomen aan de verschillen tussen leerlingen. Voorbeelden hiervan zijn het aanpassen van de groepering, het abstractieniveau van de opgaven, de instructietijd of het lestempo.

Om effectief te differentiëren is het belangrijk om te werken volgens een cyclisch proces (Bosker, 2005). Allereerst wordt de beginsituatie vastgesteld door middel van signalering en diagnose. Vervolgens wordt er gedifferentieerd en de instructie en lesstof waar nodig aangepast. Tot slot wordt er getoetst en geëvalueerd om vast te stellen of het gestelde doel behaald is.

Differentiatie kan op verschillende niveaus voorkomen (Denessen, 2017):

- *Macro-differentiatie:* Hierbij wordt het onderwijs voor verschillende leerlingen georganiseerd in een stelsel van diverse schooltypen. Voorbeelden zijn de niveaus in het voortgezet onderwijs (vmbo, havo, vwo), scholen voor speciaal onderwijs, of aparte scholen voor hoogbegaafde leerlingen.
- *Meso-differentiatie:* Dit is differentiatie tussen klassen binnen één school (externe differentiatie). Voorbeelden zijn afzonderlijke brugklassen voor vmbo, havo of vwo in een school voor voortgezet onderwijs, plusklassen of klassen voor tweetalig onderwijs. Differentiatievormen op macro- en mesoniveau zijn er vaak op gericht om de variatie in de leerlingengroep te verminderen. Hierdoor wordt het voor docenten makkelijker om onderwijs te verzorgen dat aansluit bij de leerbehoeften van alle leerlingen in een groep.
- *Binnen een klas (interne differentiatie)* (Coubergs, 2014): Coubergs maakt een onderscheid tussen convergerende en divergerende differentiatie. Bij divergerende differentiatie wordt er actief ingespeeld op verschillen tussen leerlingen, bijvoorbeeld door leerlingen homogeen te groeperen waarbij de zwakke leerlingen verlengde instructie krijgen en betere leerlingen zelfstandig gaan werken aan moeilijkere opdrachten. Een voordeel van divergeren is dat

aangesloten wordt op het niveau van leerlingen. Een nadeel van divergeren is dat zwakke leerlingen onvoldoende kunnen leren van betere leerlingen indien ze in homogene groepen worden ingedeeld. Bij convergerende differentiatie wordt heterogeniteit binnen de groep gezien als een sterkte waarbij leerlingen van elkaar kunnen leren. Hierbij worden de leerlingen bijvoorbeeld heterogeen gegroepeerd. Een voordeel van convergeren door middel van heterogene groepen is dat zwakke leerlingen kunnen leren van goede leerlingen. Een nadeel is echter dat goed presterende leerlingen minder uitgedaagd worden. Het is ook mogelijk om beide differentiatievormen te combineren (Beckers & Verstegen, 2016). De leerlingen werken hierbij in heterogeen samengestelde groepen (convergeren), waarbij de beter presterende leerlingen binnen de groep een moeilijkere rol of opdracht toebedeeld krijgen en daardoor meer uitgedaagd worden (divergeren), hetgeen de motivatie van leerlingen bevordert.

Welke beslissingen een leerkracht neemt en wat hij/zij precies doet in een klas om te differentiëren is nog onvoldoende onderzocht. Effectstudies zijn vaak gericht op de gevolgen van populaire differentiatiestrategieën, zoals het indelen van leerlingen in groepen waarbij de zwakke leerlingen verlengde instructie krijgen en de betere leerlingen zelfstandig aan de slag gaan met verdiepende opdrachten. De sleutel tot succesvolle differentiatie is echter sterk gerelateerd aan de weloverwogen instructiebeslissingen die een leerkracht neemt om het onderwijs daadwerkelijk aan te passen aan de onderwijsbehoeften van leerlingen (Deunk, Doolaard, Smalle-Jacobse, & Bosker, 2015). Recent onderzoek van Keuning et al. (2017) laat zien dat docenten die goed kunnen differentiëren doelgericht werken, verschillen tussen leerlingen goed in kaart brengen, het leerproces van de leerling continu monitoren door gebruik te maken van verschillende informatiebronnen en deze informatie doelbewust en goed doordacht gebruiken om instructies per doel aan te passen. Docenten die goed differentiëren bereiden het onderwijs voor een bepaalde periode voor waarbij ze de gegevens van de leerlingen analyseren, de doelen bepalen, de leerlingen clusteren en de aanpak bepalen om die doelen te behalen. Ze stellen ook de doelen en instructie per les vast en monitoren tijdens de les monitoren of leerlingen de stof begrijpen. Na afloop van de les evalueren de docenten hun les en stellen hun plan waar nodig bij (Keuning et al., 2017). Er wordt dus voortdurend gemonitord en bekeken welke doelen behaald moeten worden en hoe de instructie daarop afgestemd kan worden. Zijn dit bijvoorbeeld enkel de doelen gericht op het wegwerken van een achterstand? Moet er ook aandacht worden besteed aan de doelen binnen vakken waarop de leerling bovengemiddeld presteert en erg gemotiveerd voor is? Om deze beslissingen te kunnen nemen is het van belang dat de docent over veel vakinhoudelijke kennis beschikt.

De DTT biedt informatie die van waarde kan zijn bij het in kaart brengen van de verschillen tussen de leerlingen en het monitoren van de voortgang van leerlingen. Omdat de rapportage van de DTT een analyse geeft op leerling, groeps-, en schoolniveau heeft de DTT de potentie om een bijdrage te leveren aan macro-, meso-, en binnen een klas differentiatie.

7.2.2 Formatief toetsen, formatieve feedback en zelfregulerend leren

Om effectief te kunnen differentiëren dienen er voldoende data beschikbaar te zijn die inzicht geven in de prestaties en de leerbehoeften van leerlingen. Formatieve toetsen zijn een belangrijke bron van informatie om het niveau van de leerlingen te bepalen en hun ontwikkeling te monitoren.

Er bestaat geen eenduidige definitie van formatief toetsen. Verschillende concepten (b.v. classroom assessment, formative assessment en assessment for learning) worden veelvuldig in de literatuur gebruikt, maar niet consequent voor dezelfde doeleinden gehanteerd. De definities komen overeen in de omschrijving dat formatief toetsen continu plaatsvindt tijdens het onderwijs en leren, op verschillende manieren uitgevoerd kan worden, intensieve interactie tussen de docent en de leerling vereist en het leren stimuleert. Dit betekent dat toetsing niet alleen plaatsvindt na afronding van een leerproces, maar ook wordt ingezet om het leerproces te plannen en te monitoren (Sluijsmans, Joosten-ten Brinke, & Van der Vleuten, 2013). Deze elementen van plannen en monitoren zijn ook te herkennen in de differentiatie cyclus.

Belangrijke voorwaardes voor een duurzame implementatie van formatief toetsen zijn (Sluijsmans et al., 2013):

- De toetsbekwaamheid van de docenten
- Het organiseren van effectieve vormen van professionalisering
- Het stimuleren van een onderzoekende houding naar de eigen toetspraktijk
- Het creëren van een leergemeenschap waarin docenten samen reflecteren op hun werkwijze, onderzoek doen naar de relatie tussen hun manier van doceren en de resultaten van hun leerlingen en veranderingen doorvoeren ter verbetering van het onderwijs. Scholen die werken als een leergemeenschap geven sturing aan hun eigen leren waardoor zij in staat zijn zich continu te verbeteren.

Een formatieve toets is effectief voor leren als de informatie die hieruit voortkomt, bijdraagt aan het verkleinen van het verschil tussen waar een leerling zich bevindt en waar de leerling naar toe wil of moet.

Formatieve feedback is een van de bronnen van informatie die voort kan komen uit een formatieve toets. Formatieve feedback betreft de informatie voor de leerling die is bedoeld om het denken of het gedrag van de leerling zo aan te passen dat het leren wordt versterkt (Fluckiger, Vigil, Pasco, & Danielson, 2010). Deze feedback moet specifiek, simpel, beschrijvend en op de taak gericht zijn. Formatieve feedback kan gericht zijn op meerdere aspecten van een taak, zoals het product, het proces en op de vooruitgang op basis van duidelijke criteria. Duidelijke criteria en kwalitatief goede feedback zijn cruciaal voor het leerproces (Sluijsmans et al., 2013).
Effectieve feedback beantwoordt drie vragen (Hattie & Timperley, 2007):

- *Feed-up:* Waar gaat de leerling naartoe (wat zijn de beoordelingscriteria en standaarden?)?
- *Feedback:* Hoe heeft de leerling de taak uitgevoerd? Welke vooruitgang wordt geboekt ten aanzien van de beoordelingscriteria en standaarden? Wat gaat er goed en wat kan nog beter?
- *Feed-forward:* Hoe gaat de leerling nu verder (Welke aanpak is nodig om tot groei te komen)?

Het belang van feed-up, feedback en feed-forward is ook te herkennen in het model van Black en William (2009). In dit model worden hierbij drie actoren onderscheiden: de docent, de medeleerling en de leerling. Zowel, de docent, als een medeleerling kan voorzien in formatieve feedback, maar ook de leerling zelf speelt in dit model een belangrijke rol. De leerling kan zichzelf vragen stellen zoals, waar wil ik naartoe, waar sta ik nu en hoe kom ik tot de gewenste situatie? Formatief toetsen benadrukt ook de rol van de leerling als actieve deelnemer in het leerproces (Clark, 2012). Het stimuleren van de zelfregulerende vaardigheden (plannen, monitoren en toetsen) van de leerling is hierbij een belangrijk doel (Nicol & Macfarlane-Dick, 2006). Zelfregulerende leerlingen genereren meer interne feedback dan niet zelfregulerende leerlingen, reageren positiever op externe feedback

en spannen zich meer in om doelen te bereiken (Nicol & Macfarlane-Dick, 2006). Voor het proces van zelfregulerend leren is het essentieel dat leerlingen beschikken over rijke feedback en goede beoordelingsvaardigheden. Binnen het onderwijs moet dan ook ruimte worden geboden aan leerlingen om deze vaardigheden te ontwikkelen middels bijvoorbeeld reflectieve lessen, lessen waarin ze vragen kunnen stellen en toetsdialogen voeren. Zo zijn zij niet enkel afhankelijk van het oordeel van de docent. Het is van groot belang dat leerlingen van formatieve feedback worden voorzien om hun leren zelf te kunnen reguleren.

Tabel 5. Effectieve kenmerken van formatief toetsen (Black & Wiliam, 2009; Sluijsmans et al., 2013)

<table>
<tr><th></th><th>Waar werkt de leerling naar toe?</th><th>Waar staat de leerling nu?</th><th>Hoe komt de leerling naar de gewenste situatie?</th></tr>
<tr><td>Leraar</td><td rowspan="3">1. Feedback, vragen stellen, toetsdialogen rubrieken[a]

2. Feedback, reflectieve lessen, peer-assessment, rubrieken[b]

2. Self-assessment, reflectieve lessen, beoordelingsrubrieken[b]</td><td>3. Feedback, vragen stellen, toetsdialogen reflectieve lessen, beoordelingsrubrieken, summatieve toetsen[c]</td><td>4. Feedback, reflectieve lessen[d]</td></tr>
<tr><td>Medeleerling</td><td colspan="2">5. Feedback, toetsdialoog, reflectieve lessen, peer-assessment, beoordelingsrubrieken[e]</td></tr>
<tr><td>Leerling</td><td colspan="2">6. Self-assessment, reflectieve lessen, rubrieken[f]</td></tr>
</table>

[a]Welke kenmerken verhelderen voor docenten de leerdoelen en de criteria voor succes?

[b]Welke kenmerken zorgen voor begrip bij lerenden?

[c]Welke kenmerken zorgen voor effectieve discussie, taken en activiteiten in de klas die een bewijs leveren dat er geleerd wordt?

[d]Welke kenmerken zorgen dat de geleverde feedback de lerenden ook verder brengt?

[e]Welke kenmerken zorgen dat lerenden elkaar zien als een bron voor leren?

[f]Welke kenmerken zorgen dat lerenden zichzelf als eigenaar van hun eigen leerproces zien?

De formatieve feedback die voortkomt uit de DTT data is voor docenten en leerlingen waardevol om een indicatie te krijgen van waar de leerling op dit moment staat. Voor feed-forward en feed-up zijn aanvullende vormen van formatieve feedback noodzakelijk. Mogelijk kunnen hiervoor de tussendoelen worden gebruikt die zijn opgesteld als onderdeel van het DTT project.

7.2.3 Data-based decision making en opbrengstgericht werken

Om inzicht te krijgen in het leerproces van de leerling is het van belang om de voortgang van de individuele leerling te monitoren met een variatie aan toetsen die elkaar aanvullen in de informatie die ze geven over de ontwikkeling van de leerling. Maar, het is ook van belang om de data die beschikbaar zijn te analyseren op het niveau van de klas en de school. In de literatuur wordt gesproken over data-based decision making en opbrengstgericht werken.

Data-based decision making (DBDM) is het systematisch analyseren van databronnen binnen een school. Deze databronnen kunnen gestandaardiseerde toetsen, formele toetsen en klassenobservaties zijn (Hoogland et al., 2016; Van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). De resultaten van deze analyse worden gebruikt om het onderwijs, de curricula en de schoolprestaties te verbeteren. Vervolgens wordt de implementatie van deze verbeteringen geëvalueerd. DBDM vindt plaats op het niveau van de leerling, klas en school.
DBDM is een cyclisch en iteratief proces bestaand uit de volgende stappen:

- *Vaststellen doelen:* Met welk doel worden de data gebruikt ter verbetering van het leren en onderwijs
- *Data verzamelen*
- *Analyseren van de data:* Het leerproces en leerling behoeften identificeren in relatie tot de gestelde doelen
- *Interpreteren van de data* en bepalen welke acties het leren bevorderen
- *Actie ondernemen*
- *Evalueren* van de resultaten (kan leiden tot nieuwe cyclus).

Docenten vinden het moeilijk om data te gebruiken voor het verbeteren van hun onderwijs. Er zijn verschillende factoren die het gebruik van de data kunnen belemmeren of bevorderen (Schildkamp & Kuiper, 2010):

- *Toetsinstrumenten en toets proces:* Voor succesvolle implementatie moet gebruik worden gemaakt van meerdere data bronnen van hoge kwaliteit. Bovendien dienen deze data gemakkelijk toegankelijk en georganiseerd te zijn in overeenstemming met de behoeften van de gebruikers.
- *De rol van de docent:* de docent moet beschikken over voldoende kennis en vaardigheden om de data te kunnen verzamelen, analyseren en interpreteren. Bovendien moeten instructiemethodes kunnen aanpassen op basis van de data. Docenten dienen een positieve houding te hebben ten opzichte van de data, moeten het vertrouwen hebben dat ze kunnen omgaan met de data en de bereidheid tonen om te willen leren en veranderen. Het is belangrijk dat docenten, ondersteunend personeel en schoolleiders samenwerken.

- *De rol van de leerling:* Leerlingen worden in onderzoek vaak enkel omschreven als deelnemers in het DBDM proces. Er kan nog veel winst geboekt worden door hun rol verder te onderzoeken.
- *De context:* een cultuur waarin samenwerking, continue verbetering, goal setting en de vrijheid om te experimenteren wordt gestimuleerd is essentieel. Er moet bovendien sprake zijn van een duidelijke organisatiestructuur en richtlijnen voor het gebruik van data. Schoolleiders dienen doelen te stellen voor het gebruik van data en moeten de activiteiten monitoren.

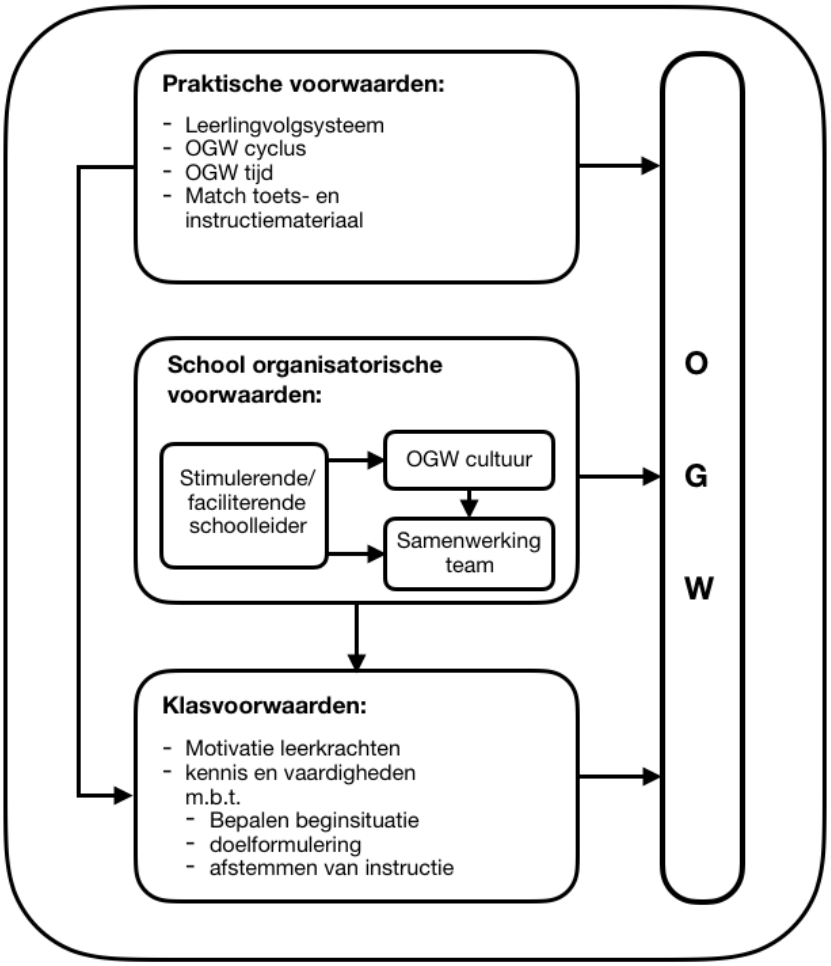
DBDM benadrukt het belang van gebruik maken van meerdere databronnen op het niveau van de individuele leerling, klas en school. De DTT kan als gestandaardiseerde toets een van de informatiebronnen zijn bij DBDM die informatie geeft op al deze niveaus.

Opbrengstgericht werken (OGW) is een onderwijscultuur waarbij systematisch en doelgericht wordt gewerkt aan het maximaliseren van prestaties voor alle leerlingen (Onderwijsinspectie, 2008). De leerlingen streven ambitieuze en expliciete doelen na en maken zo effectief mogelijk gebruik van de schaarse middelen. Voor de inrichting van het onderwijs worden instrumenten ingezet waarmee de vorderingen van leerlingen nauwkeurig gevolgd kunnen worden. Op basis van deze meetinstrumenten wordt het onderwijs beter op de uiteenlopende behoeften van leerlingen afgestemd. Het doel is om erachter te komen hoe individuele leerlingen geholpen kunnen worden en om differentiatie in de klas te ondersteunen (Heemskerk, Verbeek, Kuiper, Oomens, & Hilbink, 2014). De rol van de docent is cruciaal. OGW maakt het voor docenten mogelijk om meer gericht te zijn op de individuele leerlijn van de leerlingen. Regelmatig geven van feedback aan leerlingen op grond van de tussentijdse resultaten is een belangrijke component van OGW.
OGW verloopt volgens een evaluatieve cyclus (Visscher & Ehren, 2011):

- *Het bepalen van de uitgangssituatie van de school en lesgroepen:* In hoeverre beheersen de leerlingen de leerstof en wat zijn de leerbehoeften van individuele én groepen leerlingen.
- *Het definiëren van de gewenste situatie:* Expliciete, heldere doelen die aangeven wat het onderwijs in de toekomst moet opleveren. Dit moet doelgericht i.p.v. activiteitengericht gedrag stimuleren.
- *Bepalen van de instructiebenadering:* Dit gebeurt op basis van kennis over de beoogde en beginsituatie van de leerlingen en de groep. Er worden aanpassingen gedaan aan zowel de leerstofinhoud, didactiek als klassenmanagement.

Door het doorlopen van deze cyclus kunnen de aandachtspunten binnen het leerproces van de leerling en de groep worden geïdentificeerd. Vervolgens kan gericht worden gewerkt aan verbetering. Figuur 1 geeft de voorwaarden weer die binnen scholen belangrijk zijn voor succesvol OGW.

Figuur 26. Voorwaarden voor succesvolle implementatie van OGW (Visscher & Ehren, 2011)



De eerste stap van de evaluatieve cyclus van OGW is het bepalen van de uitgangssituatie van de school en de lesgroepen. De DTT geeft voor de vakken Nederlands, Engels en Wiskunde per subdomein aan wat de leerbehoeftes van de individuele leerlingen en groepen leerlingen zijn. De DTT kan dus input leveren voor deze eerste stap van OGW.

Kortom, de DDT data kunnen gebruikt worden als een bron van informatie bij data-based decision making en opbrengstgericht werken, zowel op het niveau van de individuele leerling, als van de klas en van de school.

7.3 Aanbevelingen

Op basis van de literatuur zijn er hieronder enkele aanbevelingen geformuleerd die kunnen bijdragen aan een verdere benutting van de informatie uit de DTT voor docenten en leerlingen in het bijsturen van het leerproces van de individuele leerling.

7.3.1 Aanbevelingen voor diagnostisch en formatief toetsen

- **Cyclische in plaats van eenmalige afname**

De kern van differentiëren is de herhaaldelijke cyclus waarbij doelen worden gesteld, het niveau van de leerling wordt bepaald en tot slot wordt gekeken hoe de gewenste situatie kan worden bereikt. De DTT is een geschikt instrument om het niveau van de leerling te bepalen. Om de voortgang te kunnen blijven monitoren is het aan te bevelen om de toets niet eenmalig af te nemen, maar meerdere malen.

- **DTT als onderdeel van het leerlingvolgsysteem**
  Om een goed beeld te krijgen van het niveau en ontwikkeling van de leerlingen is het belangrijk om gebruik te maken van een verscheidenheid aan toetsinstrumenten die op meerdere momenten worden afgenomen. De DTT is uitermate geschikt om inzicht te bieden in het huidige niveau van de leerlingen voor de vakken Nederlands, Engels en Wiskunde. Vooraf moet bepaald worden hoe de DTT van aanvulling is op de huidige toetsen en op welke manier de DTT onderdeel gaat uitmaken van het leerlingvolgsysteem.
- **Flexibele afname op onderdelen in plaats van gehele afname**
  Op dit moment is het afnemen van de DTT erg tijdsintensief (maximaal 3 uur per vak). De lange duur van de toets heeft een negatief effect op de motivatie van de leerlingen. Bovendien is het moeilijk voor scholen om de toets in te plannen. Docenten zijn in staat om op basis van observaties en toetsdata een inschatting te maken van welke onderdelen van hun vak boven niveau of voldoende beheerst worden door de leerling en welke onderdelen nog aandacht nodig hebben. Door de docent inspraak te geven in welke onderdelen van de toets afgenomen moeten worden kan de toets ingekort worden. Hierdoor sluit de toets ook beter aan bij de leerbehoefte van de leerling.

7.3.2 Communiceren van uitkomsten van diagnostische toetsen met docenten en leerlingen

- **Aanvullen van de rapportage met narratieve informatie:**
  Voor zowel leerlingen als docenten is het van belang om de DTT rapportages aan te vullen met narratieve feedback. Uit de eindmeting van de pilot DTT kwam naar voren dat docenten moeite hebben met het interpreteren van de DTT data (van der Wel, Paulussen-Hoogeboom, & Dekker, 2017). Daarom is een uitleg van de interpretatie van de scores van belang. Hierdoor kunnen docenten een inschatting maken van het huidige groeps- en individuele niveau en bepalen welke verdere stappen ondernomen moeten worden om de leerdoelen van de groep en de individuele leerlingen te bereiken. De narratieve informatie zou vooral informatie moeten bevatten over aspecten die ondermaats of uitzonderlijk goed scoren zodat docenten en leerlingen gericht kunnen werken aan een plan van aanpak.
  Tijdens de pilot zijn al enkele pogingen ondernomen om narratieve informatie toe te voegen aan de DTT rapportage. Deze uitleg bood de docenten echter niet de informatie waar zij behoefte aan hadden. Docenten gaven aan graag concreet aangereikt te willen krijgen wat zij in hun instructie of begeleiding van de leerlingen moeten veranderen. Nader onderzoek naar hoe op een juiste manier invulling gegeven kan worden aan de informatiebehoefte van de docent is noodzakelijk.
- **Creëren van een professionele leergemeenschap:**
  Het creëren van een leergemeenschap een belangrijke voorwaarde voor het implementeren van formatief toetsen. Binnen de scholen moet een cultuur worden gestimuleerd waarin docenten samen reflecteren op hun werkwijze, de prestaties van hun leerlingen en gezamenlijk ideeën genereren over hoe veranderingen in instructies en begeleiding het

leerproces kunnen bevorderen. De informatie uit de DTT data kan belangrijke input leveren voor deze discussie en reflectie.

- **Professionaliseren van docenten:**
  Er moet aandacht worden besteed aan hoe de docenten begeleid worden in het omgaan met en interpreteren van de DTT data. Op leerling niveau biedt de DTT een gedetailleerd overzicht van de sterke en zwakke punten van de leerling. De DTT groepsdata kunnen de docenten helpen bij zowel het onderscheiden van subgroepen en het aan de hand hiervan aanpassen van het groepsplan, als bij het identificeren van de leerbehoeften van individuele leerlingen. Professionalisering van docenten is hierbij cruciaal. Docenten moeten getraind worden in hoe ze met behulp van de DTT data de leerling behoeften kunnen identificeren en het groepsplan kunnen aanpassen.
- **Dialoog aangaan met de leerling**
  Zowel op groeps-, als leerling niveau is het belangrijk dat de cyclus van feed-up, feedback en feed-forward wordt voltooid. Voorafgaand aan de afname van de DTT dienen zowel de docent als de leerling een duidelijk beeld hebben van de beoogde doelen. Na afname van de DTT moeten de docent en leerling bekijken waar de leerling nu staat en hoe de leerling naar de gewenste situatie komt ofwel welke aanpassing er nodig is in de instructie. Op deze manierkan de DTT daadwerkelijk bijdragen aan het leerproces. Enkel het geven van feedback is geen garantie voor gedragsverandering van de leerling. Voor effectieve zelfregulatie is het van belang dat de leerling inzicht heeft in de huidige situatie én weet welke stappen moeten worden ondernomen om tot de gewenste situatie te komen. Het is belangrijk dat de leerlingen leren hoe ze fouten kunnen herkennen en weten hoe ze deze aan moeten pakken. Dit kan worden bereikt door leerlingen uit te dagen op hun eigen leren te reflecteren. Dit reflectieproces kan worden bevorderd door de dialoog met de docent en medeleerlingen aan te gaan die de leerling helpt om het eigen leerproces vorm te geven. Deze dialoog draagt ook bij aan het krijgen van inzicht in de beoordelingscriteria, wat er verwacht of beoogd wordt van de leerling.
- **Koppeling van toetsresultaten aan opdrachten**
  Een belangrijk onderdeel van differentiëren is het aanbieden van een vervolgstap waardoor een leerling alsnog de gestelde doelen kan halen. Het is aan te raden om in de toekomst de uitslag van de toets meteen te koppelen aan modules met verschillende opdrachten. De leerling kan dan zelfstandig deze opdrachten selecteren en inzetten voor verbetering van de eigen aandachtspunten. Bovendien is het voor de docent ook makkelijker om te differentiëren wanneer er leerstof wordt aangereikt op basis van de specifieke scores van de leerlingen (Keuning et al., 2017).

8 Conclusies

In dit hoofdstuk bespreken we de conclusies vanuit het voorliggende onderzoek naar de benutting van de DTT in termen van de geformuleerde onderzoeksvragen.

1. Welke rol kan het gebruik van de DTT in opbrengstgericht werken op klas-, afdelings- en schoolniveau spelen en in welke mate verschilt deze bruikbaarheid tussen klassen, afdelingen en scholen?

Hoewel de DTT is ontworpen vanuit leerlingperspectief kunnen de resultaten ook aangewend worden om op schoolniveau diagnoses te stellen. Door het ontwerp van een toets kunnen de meetfout en lengte van de toets worden beïnvloedt. Afhankelijk van het doel van de toets zal deze afweging verschillen. Een optimaal ontwerp voor een toets die scores op leerlingniveau moet bepalen is anders dan een optimaal ontwerp voor een toets voor scores op schoolniveau. Om een toets meerdere functies te geven kan bij het ontwerp echter ook een afweging gemaakt worden tussen de verschillende doelen. In dit rapport hebben we een continue indicator geconstrueerd die geaggregeerd kan worden op klas-, afdelings- en schoolniveau, waarmee een meer precieze maat wordt gecreëerd dan wanneer de categoriale scores van de DTT worden geaggregeerd. De schoolscores op de verschillende hoofdaspecten en deelaspecten van de DTT geven klassen, afdelingen en scholen precieze indicaties van hoe de prestaties van leerlingen zich verhouden tot een landelijke standaard, of hoe deze onderdelen verschillen ten opzichte van het overkoepelende niveau voor dat schoolvak. Op deze manier kunnen zowel absolute als relatieve sterktes van klas, afdeling en school in beeld gebracht worden, en kan daar vervolgens op bijgestuurd worden. Door bij de selectie van vragen bij een toets rekening te houden met het belang om naast diagnoses per leerling ook op geaggregeerd niveau informatie te verkrijgen, kan de betekenis van de toets op dit niveau nog verder worden vergroot.

De verschillende aggregaties zullen op verschillende niveaus nuttig zijn. Leraren zullen met name behoefte hebben aan rapportages op klasniveau, afdelingshoofden aan rapportages op afdelingsniveau en schoolleiders aan rapportages op schoolniveau. De uitsplitsing van de DTT-scores naar deelaspecten zal vooral nuttig zijn voor leraren in de klasrapportages, wat ook bevestigd wordt door schoolleiders en leraren zelf in de schoolbezoeken. Maar ook voor afdelingshoofden en schoolleiders is het informatief om, bijvoorbeeld, te weten dat veel leerlingen op de school of afdeling tekort schieten op het spellen van werkwoorden. Dit zal ook afhangen van hoe scores verspreid zijn over de verschillende klassen. Wanneer alle klassen op de school op het deelaspect tekort schieten is dit een duidelijk signaal om dit schoolbreed aan te pakken en wellicht methodes aan te passen. Wanneer het vooral in bepaalde klassen is geconcentreerd, dan lijkt bijsturing op het niveau van klas of leraar nuttiger. In het algemeen lijken de overkoepelende scores en de hoofdaspecten vooral informatief om de belangrijkste pijnpunten te signaleren, waarna vervolgens ingezoomd kan worden op de onderliggende deelaspecten om te analyseren waar er vooral tekortkomingen zitten.

2. Hoe kijken de verschillende belanghebbenden in het onderwijsveld naar de betekenis van de DTT en welke consequenties heeft dit voor het gebruik van de DTT?

Perspectief bestuur

Vanuit dit perspectief vindt men het interessant dat de DTT in kaart brengt hoe de scholen binnen het bestuur ervoor staan ten opzichte van de landelijke norm of het gemiddelde van andere scholen. Eén van de bevraagde scholen geeft aan dat het voor het bestuur vooral interessant zou zijn om te zien hoe de school het doet ten opzichte van scholen in de nabije regio. Als blijkt dat een school beter presteert ten opzichte van andere scholen in de nabije regio, dan zou dat een positieve invloed kunnen hebben op de leerlingaanmeldingen. Een vergelijking met scholen in de nabije regio was tijdens de pilot DTT niet per definitie mogelijk, aangezien de DTT schooleigen, niet-openbare informatie betrof. Alleen wanneer scholen dit zelf wilden konden zij onderling rapportages met elkaar vergelijken.

Perspectief directie
De directie van een school wil allereerst weten hoe de school ervoor staat ten opzichte van andere scholen en ten opzichte van een inhoudelijk bepaalde norm. Voor dit doel wordt de schoolrapportage van de DTT resultaten op hoofdaspecten van de vaardigheid veelal voldoende bevonden. Daarnaast speelt de schoolrapportage een rol in het in kaart brengen van de aandachtspunten die er liggen binnen de kernvakken. De schoolleiding kan aan de hand van de schoolrapportage het gesprek op gang brengen binnen de secties over waar sectiebreed de knelpunten liggen. Daarmee vormt de schoolrapportage een gevalideerd middel om probleemgebieden onder de aandacht te brengen en discussies op gang te brengen. Voor dit doel is het voor de scholen wel wenselijk dat de schoolrapportage meer detailniveau bevat dan alleen de hoofdaspecten van de vaardigheid.

Perspectief afdeling
Vanuit het perspectief van de afdeling is opnieuw de schoolrapportage van de DTT interessant om te zien waar het goed gaat en waar minder goed. Ook kunnen verschillende klas- en groepsrapportages binnen de sectie vergeleken worden om inzicht te krijgen in waar de verschillen en overeenkomsten zitten. De sectie kan aan de hand van de schoolrapportage de resultaten duiden in relatie tot schoolrapportages van andere onderwijsstromingen binnen de school. De DTT zorgt zo voor grip en sturing door inzichtelijk te maken waar schoolbreed de aandachtspunten zitten die aangepakt moeten worden.

Perspectief docent
Voor de docent is de rapportage op klas- of groepsniveau het meest waardevol, omdat hij of zij op de verschillende onderdelen uit de rapportage kan zien waar voor welke groepen leerlingen verbeterpunten en sterke punten liggen. Uit de DTT-rapportage kwam volgens de bevraagde docenten echter onvoldoende naar voren hoe de docent vervolgens aan de slag kan gaan met de rapportage om het onderwijs te verbeteren (zie onderzoeksvraag 3). Duidelijke vervolgstappen bij de onderdelen van de rapportage zijn voor de docenten cruciaal, willen deze bruikbaar zijn voor hen om het leren van leerlingen verder te brengen.

Perspectief leerling
Hoewel de UM tijdens de consultatieronde bij scholen geen leerlingen heeft gesproken, ging het gesprek vaak over de bruikbaarheid van een instrument als de DTT op leerlingniveau. In een tijd waarin steeds meer aandacht is voor individuele leerroutes en formatief toetsen, is het voor zowel leerlingen als hun ouders prettig om tussentijds te zien hoe het gaat, op basis van een landelijk

genormeerd instrument als de DTT. Voor de docent blijkt het in de praktijk lastig om met elke leerling aan de slag te gaan met de rapportage, aangezien hij daarvoor te veel leerlingen onder zijn hoede heeft. Op leerlingniveau is er met de DTT tijdens de pilot daarom niet zoveel gedaan als wenselijk zou zijn. Uit gesprekken met scholen blijkt dat docenten vooral aan de slag gaan met die leerlingen waar de grootste achterstanden en knelpunten te zien zijn.

Behoefte aan methode-onafhankelijk instrument
De meeste scholen geven aan dat een instrument als de DTT een waardevolle rol zou kunnen spelen in de determinatie van leerlingen aan het eind van de onderbouw, richting de bovenbouw. Een dergelijk diagnostisch instrument zou dan een methode-onafhankelijk, landelijk genormeerd gegeven zijn naast andere gegevens. De bruikbaarheid van de DTT als vergelijking met de resultaten uit andere instrumenten werd meermaals aangehaald door scholen. Het kan de school informatie geven over de vraag of de lat binnen andere instrumenten in de school wel op de juiste hoogte ligt. Overigens bleek uit de gesprekken met scholen dat de DTT resultaten (soms flink) afweken van wat er uit andere instrumenten naar voren komt. Het vermoeden bestaat dat dit ook te maken heeft met docent- en leerlingmotivatie, moeilijkheidsgraad (vooral wat betreft de wiskundetoets) en de lange duur van de toets tijdens de DTT pilot (zie ook onderzoeksvraag 3).

3. Onder welke voorwaarden zullen de verschillende belanghebbenden met diagnostische en formatieve toetsen willen werken?

In vraag 2 is aan de orde gekomen wat de DTT op verschillende niveaus binnen de school *kan* betekenen. Uit gesprekken van de UM met Limburgse scholen die aan de DTT pilot hebben deelgenomen blijkt dat de DTT niet altijd deze rol heeft vervuld op de verschillende niveaus binnen de scholen. Scholen hebben enkele factoren benoemd die hierin hebben meegespeeld. Bovendien is meegedacht over de vraag waar diagnostische instrumenten in de toekomst aan moeten voldoen, willen zij van toegevoegde waarde zijn in het leveren van maatwerk om het onderwijs te verbeteren. Onderstaand zijn de belangrijkste punten op een rij gezet.

Organisatie van de afname
De DTT afname zorgde voor een flinke organisatiedruk op scholen. Met name vanwege de grote toetsduur per vak kwam er nogal wat kijken bij het inplannen van de toets (en uitval van lessen) en de beschikbaarheid van computerlokalen. Scholen adviseren dan ook om de afname van diagnostische instrumenten in de toekomst laagdrempeliger te maken in de organisatie, door bijvoorbeeld de afname mogelijk te maken via een draadloze internetverbinding, de afnameperiode uit te breiden en het genereren van de rapportages te vereenvoudigen (de rapportages moesten nu één voor één gedownload worden, het zou praktischer zijn wanneer alle rapportages in één keer te downloaden zijn).

Eenvoudige, duidelijke rapportages inclusief vervolgstappen
Scholen hebben vooral behoefte aan eenvoudig te interpreteren rapportages. Ze geven aan dat het na het zien van de rapportage voor de docent (en leerling) duidelijk moet zijn waar, maar vooral ook hoe, er aan bepaalde punten gewerkt kan worden. Met name het 'hoe nu verder' was voor scholen onvoldoende duidelijk uit de DTT rapportage. Docenten zouden vooral graag zien dat het instrument hen direct faciliteert in het toepassen van differentiatie in de klas. Sommige scholen gaven aan dat

het prettig zou zijn wanneer het instrument automatisch leerlingen indeelt in groepen, waarbij het per groep direct duidelijk is welke vervolgstappen (bijvoorbeeld: instructies, lesmethodes, opdrachten) mogelijk zijn. Zo kan de docent zich meer richten op het onderwijzen en begeleiden van leerlingen.

Compleet meetsysteem voor de kernvakken
Uit de gesprekken met scholen blijkt dat er behoefte is aan een compleet meetsysteem voor de kernvakken in met name de onderbouw, dat informatie op leerling-, groeps- en schoolniveau geeft. Bovendien is het van belang dat er een methode-onafhankelijk, genormeerd instrument is (zoals de DTT), als extra gegeven naast andere bronnen waaruit informatie over de leerlingen gehaald wordt. Scholen zouden ook graag zien dat diagnostische instrumenten gekoppeld kunnen worden aan andere instrumenten (zoals methodes, LVS of andere methode-onafhankelijke bronnen) zodat passende vervolgacties meteen voor handen liggen.

Overlap met andere instrumenten
De toets- en monitordruk op scholen is hoog. Scholen geven aan kritischer te gaan kijken aan welke toetsen en vragenlijsten ze deelnemen. Veel instrumenten kennen namelijk een overlap in te meten thema's. Het is belangrijk dat diagnostische instrumenten toegevoegde waarde hebben ten opzichte van wat scholen al in huis hebben.

Adaptief en flexibel inzetbaar instrument
Een adaptief en flexibel inzetbaar instrument zou ten eerste een oplossing kunnen bieden voor de grote toetsduur waar veel scholen tegenaan liepen. Organisatorisch was het lastig om voor zoveel leerlingen een toets van drie uur in te plannen in het lesrooster (met een grote belasting op het gebruik van de computerruimtes). Ook bleek het voor leerlingen, vooral in het vmbo, zwaar om zo lang geconcentreerd en gemotiveerd te werken aan de toets.[35] Een adaptief en flexibel inzetbaar instrument zou ook beter aansluiten bij individuele leerroutes. De scholen hebben behoefte aan een instrument dat in de hele onderbouw flexibel inzetbaar is op verschillende momenten en in delen afneembaar is. Op die manier kan er, zelfs per leerling, een keuze gemaakt worden in welke onderdelen wanneer getoetst worden. Bovendien kunnen leerlingen sommige onderdelen op een later moment nog eens opnieuw maken, om te zien of er sprake is van vooruitgang.

Brede steun van het instrument is cruciaal
Het is van belang dat het instrument omarmd wordt door alle lagen in een school. Het besluit tot deelname aan de pilot DTT was vaak genomen door één persoon in de organisatie (vooral op directieniveau), terwijl alle betrokkenen nodig zijn voor een succesvol gebruik van het instrument. Om dit te bereiken zouden scholen aan de voorkant nog meer bij de ontwikkeling van het instrument betrokken kunnen worden.

4. Is de afweging die (impliciet) gemaakt is voor de DTT tussen verschillende aspecten van de toets, ook voor docenten het meest optimaal bij het gebruik van de toets op leerlingniveau?

[35] In de pilot bestonden er wel mogelijkheden om de toets in delen af te nemen, maar daar is geen gebruik van gemaakt.

Aan scholen zijn de leerlingmodellen van de DTT Nederlands de DTT wiskunde voorgelegd. Bij het onderzoeken van de gewenste breedte en diepte van de te toetsen vaardigheden is volgens de scholen een juiste balans tussen algemeen en specifiek van belang. Bij een te algemeen niveau van rapporteren zal de toets onvoldoende bruikbaar zijn om het leren verder te brengen. Een teveel aan data en informatie zal echter ook leiden tot onvoldoende bruikbaarheid, omdat docenten en leerlingen niet goed meer weten waar te beginnen.

Bij het leerlingmodel van de DTT Nederlands geven de meeste scholen aan dat de huidige indeling in hoofd- en deelaspecten goed werkbaar is voor het gebruik van het instrument op leerlingniveau. Een enkeling geeft aan dat eigenlijk nog meer diepte gewenst is. Een rapportage op alleen de hoofdaspecten zou in elk geval te algemeen zijn. Over het leerlingmodel DTT wiskunde is wat meer verdeeldheid. Sommigen geven aan dat een rapportage op domeinniveau voldoende is, aangezien de meeste methodes een zelfde indeling hanteren en de DTT daar mooi op aansluit. Anderen geven aan dat een rapportage op alleen domeinniveau weinig toegevoegde waarde heeft. Docenten weten vaak al goed hoe hun leerlingen het doen op de domeinen. Bovendien komt dit type informatie ook naar voren uit andere instrumenten, waarvan scholen gebruik maken.

Een uitbreiding van van diagnostische informatie over andere vakken acht men niet nodig. Nederlands, Engels en wiskunde zijn niet voor niets kernvakken, deze hebben ook hun weerslag op de andere vakken. Naast schrijfvaardigheid zou men het toetsen van de leesvaardigheid wel een interessante uitbreiding vinden.[36] Leesvaardigheid vormt onderdeel van het eindexamen. Bovendien is in andere vakken, zoals rekenen, een goede leesvaardigheid van belang. Het zou voor scholen daarom zeer interessant zijn als leesvaardigheid niet alleen gekoppeld is aan het vak Nederlands, maar vakoverstijgend getoetst zou kunnen worden.

5. Hoe kunnen uitkomsten van de DTT worden gerelateerd aan andere gegevens over de achtergrond en ontwikkeling van leerlingen om daarmee een zo bruikbaar mogelijke rapportage te maken, voor datagericht/opbrengstgericht werken op de school en t.b.v. de ingezette beleidsplannen van school en bestuur?

Voor scholen die deelnemen aan de OnderwijsMonitor Limburg kunnen de resultaten uit de DTT gerelateerd worden aan achtergrondgegevens en eerdere toetsscores. In het rapport zijn enkele voorbeelden gegeven van toepassingen, zoals een koppeling met gegevens van de Centrale Eindtoets PO en de samenhang tussen de prestaties op de verschillende deelaspecten van de DTT (gemeten op basis van de continue indicator) en de sociaaleconomische status van de leerling. Hier komt uit naar voren dat er duidelijk een relatie is tussen prestaties op de DTT en prestaties op de Centrale Eindtoets, maar ook dat specifieke deelaspecten van de DTT vaak niet het sterkst correleren met de onderdelen op de Centrale Eindtoets die het meest vergelijkbaar zijn. Dit suggereert dat de prestaties op deelaspecten veranderlijk zijn door de tijd. Verder vinden we dat de relatie tussen prestaties op de DTT en de achtergrond van leerlingen kunnen verschillen tussen de verschillende deelaspecten van de toets.

[36] Leesvaardigheid was binnen de DTT pilot wel in ontwikkeling voor de vakken Engels en Nederlands.

De koppeling die gemaakt is, is waardevol om meer context te geven aan de resultaten van de toets. Het laat bijvoorbeeld toe om te concluderen of achterstanden zijn ontstaan in de onderbouw of al eerder, en in hoeverre deze terug te herleiden zijn op de leerlingpopulatie of eerder op specifiek schoolbeleid. Deze informatie kan voor zowel leerling als school nuttig zijn, om het leerproces formatief bij te sturen. Ook kan de koppeling direct gebruikt worden om schoolrapportages te 'corrigeren' voor SES of voor prestaties op de Centrale Eindtoets PO.

De meeste bevraagde scholen vinden het interessant om rapportages op schoolniveau te ontvangen waarin de uitkomsten van een instrument zoals de DTT gerelateerd worden aan leerlingkenmerken en achtergronden. Met goede data kan men de kwaliteit van het onderwijs op school bewaken en sturen op verbeteringen. Soms is er binnen de school wel een vermoeden waar een achterstand binnen een bepaald vak vandaan komt, maar is er onvoldoende bewijs in handen om interventies erop te zetten. Door verschillende verbanden in beeld te hebben, ook in relatie tot andere scholen, kunnen vermoedens getest worden en andere verbanden zichtbaar worden. Met de data in handen kan de directie discussies op gang brengen in de sectie en kunnen beleidsinterventies ontwikkeld worden.

Enkele scholen geven aan dat het ook richting het schoolbestuur prettig is als aangetoond kan worden hoe de resultaten samenhangen met leerlingkenmerken en achtergronden. Het aannamebeleid ten aanzien van LWOO-leerlingen zal bijvoorbeeld zichtbaar zijn in de resultaten. Een correctie geeft het meest eerlijke beeld van hoe de school ervoor staat.

Andere scholen geven echter aan niet direct behoefte te hebben aan meer rapportages op schoolniveau over toetsscores in relatie tot leerlingkenmerken en achtergrondgegevens. Er ontstaat het risico op een te veel aan data. Meer maatwerk in de rapportages zou wenselijk zijn, waarbij deze vooral data aan de school geven die passen binnen de eigen aandachtsgebieden en kwaliteitsdoelen. Verder geeft deze groep scholen aan dat het perspectief meer bij de leerling zou moeten liggen. De rapportage zou volgens hen meer gefocust moeten zijn aan dat wat daadwerkelijk bijdraagt aan het leren van de individuele leerling.

6. Hoe kunnen de individuele items die worden verkregen bij de afname van de DTT het beste worden geaggregeerd om doelmatig informatie op school- en stelselniveau te genereren?

Uit onze analyses blijkt dat er op leerlingniveau veel onzekerheid is in de diagnoses op deelaspect, en dat de uitsplitsing in de DTT pilot vanuit dat perspectief te ver lijkt doorgevoerd, voor in ieder geval de vakken Nederlands en wiskunde. Dit wordt ook bevestigd door het feit dat voor de Engels toets, die een minder brede uitsplitsing heeft, er duidelijk meer precisie is in de uitkomsten op deelniveau. Om op leerlingniveau per deelgebied zinvolle feedback te geven is een vergroting van de nauwkeurigheid van de toets vereist. Bij de ontwikkeling van de DTT is om deze reden al gekozen voor een latent-klasse-model, waarbij alleen toetsresultaten in de vorm van 'onder niveau', 'op niveau', of 'boven niveau' worden berekend. Hoewel deze aanpak psychometrisch gezien ook bepaalde voordelen heeft vergeleken met een continue schaal, lijken onze analyses erop te wijzen dat deze in de specifieke context van de DTT niet sterk doorwegen. Nadeel van de scores in de vorm van deze drie klasses is dat aan de ene kant het verschil tussen twee leerlingen die net wel of net

niet onder niveau scoren groter lijkt dan het is en dat omgekeerd er geen duidelijke informatie beschikbaar is over hoe ver een leerling onder niveau scoort.

Op schoolniveau is de precisie van de toets veel groter. Dat betekent dat de huidige uitsplitsing naar deelaspecten wel precieze schattingen op schoolniveau toelaat en dat feedback over hoe de school het doet op deze deelniveaus dus zeer precies en informatief is. Op dit aggregatieniveau verdwijnt dan ook een eventueel voordeel van een uitslag in klassen. Voor schoolfeedback ligt een continue maat veel meer voor de hand, omdat deze ook verschillen binnen categorieën meeneemt. Dit geldt ook voor het stelselniveau. Vergelijkingen laten zien dat de meer precieze continue indicator tot heel andere conclusies kan leiden over de relatieve prestaties van een klas of school, vergeleken met het aggregeren van de categoriale indicator. Het meenemen van verschillen *binnen* de categorieën onder, op of boven niveau kunnen dus veel uitmaken voor conclusies over het niveau van een school op de verschillende aspecten van de toets, of op het overkoepelende vak als geheel. De beste wijze van aggregatie is in dit geval dus om eerst op leerlingniveau een continue vaardigheidsscore te construeren en deze vervolgens te aggregeren op klas-, school-, of stelselniveau.

Uit de schoolgesprekken komt verder ook naar voren dat de toegevoegde waarde van de DTT ten opzichte van andere instrumenten vooral zit in die diagnoses op deelaspect. Deze diagnoses bieden docenten concrete handvatten om mee te werken. De vraag is er dus vanuit de scholen, en het bereiken van een voldoende mate van precisie op dit niveau van de toets is dus een belangrijke aandachtspunt in de verdere ontwikkeling van deze instrumenten.

7. Zijn er aanpassingen in de opzet van de toets denkbaar die de informatieve waarde van de toets op school- en stelselniveau vergroten, zonder daarbij veel inbreuk te doen aan de waarde op leerlingniveau?

De DTT pilot is opgezet om op het niveau van de leerling de meest informatieve diagnoses te stellen. Dit kan spanningen opleveren met het gebruik van de DTT gegevens voor scores op klas- of schoolniveaus, zoals in dit rapport gedaan is. Daarnaast is de opzet van de DTT gebeurd met het oog op het latente klassemodel dat gebruikt is om categoriale uitkomsten te construeren, terwijl er in dit rapport met name gebruik is gemaakt van een continue indicator. Analyses laten echter zien dat dit spanningsveld beperkt is. Zo zien we bijvoorbeeld dat het toevoegen van vragen die specifiek onderscheidend zijn rond de twee grenzen van het latente klassemodel geen sterke afbreuk doen aan de algehele precisie van de continue score. Hierbij moet wel aangetekend worden dat dit geldt gegeven de huidige kwaliteit van de items. Wanneer de itembanken van diagnostische toetsen aangevuld en rijker worden, dan kunnen deze afwegingen een sterkere rol spelen. In dat geval kunnen methodes toegepast worden die het gebruik op leerlingenniveau en het gebruik op school- of stelselniveau gezamenlijk optimaliseren, daarbij een combinatie van items selecterend die voor beide doelen nuttig is.

8. Welke rol kan diagnostisch en formatief toetsen spelen op stelselniveau?

De informatie uit een toets als de DTT kan ook gebruikt worden om het niveau van leerlingen op het niveau van het stelsel (bijvoorbeeld vmbo-kb, vwo) te analyseren. Door leerlingen door de tijd heen te vergelijken, kunnen conclusies worden getrokken over de mogelijke verbeteringen of

verslechteringen binnen het onderwijssysteem als geheel. Op dit moment kunnen dit soort vergelijkingen alleen gemaakt worden aan het eind van groep 8 (Centrale Eindtoets) en aan het eind van de middelbare school (eindexamen). Daar tussen ligt een lange periode waarin een dergelijk vergelijkbaar beeld ontbreekt. Op basis van de DTT pilot zijn dit soort vergelijkingen nog lastig, omdat de toets maar drie afnamejaren heeft gehad waarin er logischerwijs veel veranderingen zijn doorgevoerd, en waarin ook de samenstelling van de deelnemende scholen sterk wisselend is geweest. In de ontwikkeling en het verdere gebruik van diagnostisch en formatief toetsen zou er meer aandacht besteed kunnen worden aan dit nut van dergelijke instrumenten, waarbij bijvoorbeeld voor gehele regio's of voor het totale Nederlandse onderwijssysteem bekeken kan worden hoe het zit met de ontwikkeling van bepaalde specifieke deelaspecten door de jaren heen. Zeker in relatie tot de gespecificeerde kerndoelen kan dit zeer waardevol zijn.

Appendix 1 Overkoepelende onderzoeksvragen

Voor het DTT project zijn vier overkoepelende onderzoeksvragen geformuleerd, die beantwoord worden vanuit het onderzoek aan zowel de Universiteit van Maastricht (UM) als vanuit het onderzoek aan de Universiteit Twente (UT).

1) Voor welke scholen, afdelingen, klassen en groepen leerlingen heeft de DTT de meeste relevantie en bruikbaarheid?

In het kleinschalige deelonderzoek naar de DTT vanuit de UT hebben drie katholieke voortgezet onderwijs scholen, twee algemeen bijzonder voortgezet onderwijs scholen en één openbare voortgezet onderwijs school deelgenomen. Zij hebben allen het 'basistoezicht' als beoordeling van de Inspectie van het onderwijs. Het aantal leerlingen op deze scholen varieert tussen de 500 en 1200 leerlingen. Van deze scholen zijn twaalf docenten en twaalf leerlingen geïnterviewd. In tabel 2 en 3 zijn de kenmerken gerapporteerd. Veel van deze docenten en leerlingen hebben de DTT in juni 2017 nog niet gebruikt.

In het onderzoek van de UM heeft er een kleinschalige gespreksronde plaatsgevonden met negen schoollocaties in Limburg (waarbij 12 schoollocaties vertegenwoordigd werden). Het gaat om locaties in midden Limburg en zuid Limburg. De gesprekken werden meestal gevoerd met een sector/locatiedirecteur. In sommige gevallen was er ook een kwaliteitsmedewerker, docent of beleidsadviseur bij. De meeste schoollocaties hebben tijdens de pilot aan één of meer afnames van de DTT deelgenomen. Vijf schoollocaties waren niet betrokken bij de pilot.

Het is lastig om binnen dit deelonderzoek van de UM een onderscheid te maken in bruikbaarheid van de DTT tussen (type) scholen, afdelingen, klassen en groepen leerlingen. Dit gelet op het aantal scholen waarmee gesproken is en het feit dat er vooral met leden op directieniveau gesprekken gevoerd zijn. Ook uit de gesprekken die de UM heeft gevoerd met scholen die deelnamen aan de DTT blijkt in het algemeen naar indruk van de betrokkenen in de school dat er onvoldoende gedaan wordt met de DTT rapportages.

Volgens de docenten en leerlingen uit zowel het onderzoek van UM als UT speelden een aantal factoren een rol bij het niet of nauwelijks gebruiken van de DTT rapportages. Deze worden beschreven als antwoord op de tweede overkoepelende onderzoeksvraag.

2) Welke factoren zijn hiervoor bepalend?

Uit het kleinschalige deelonderzoek naar de DTT rapportages vanuit de UT blijkt dat de meeste docenten en leerlingen de DTT rapportages nog niet hadden gezien toen het interview plaatsvond (4 maanden nadat de rapportages beschikbaar waren in de school). Dat is een van de factoren voor het niet hebben gebruikt van deze rapportages voor het nemen van een vervolgstap. Een andere factor beschreven door sommige docenten en een leerling is dat de rapportages niet op item-niveau laten zien of en welke fouten er zijn gemaakt door de leerlingen. Met de diagnose op hoofdaspecten/domeinen hebben de docent en leerling veel minder goed zicht op waar een leerling staat. Wat ging er bijvoorbeeld precies fout bij het deelaspect "correcte zinsbouw hanteren" bij het

hoofdaspect “woord- en zinsniveau”? Een andere factor beschreven door sommige docenten en een leerling is dat de DTT rapportages geen adviezen geven over wat de vervolgstap van de docent of leerling moet zijn en dat zij bepaalde kennis en vaardigheden missen om dit zelf te bepalen. Nog een andere factor beschreven door sommige docenten is dat de werkdruk te hoog is om data te gebruiken om hun onderwijs te verbeteren. Sommige anderen docenten benoemden factoren zoals gebrek aan heldere terminologie in de DTT rapportages, gebrek aan vertrouwen in de data uit de DTT rapportages (omdat deze niet altijd overeenkomen met hun voorkennis van de leerlingen) en geen verplichting door de schoolleider om de rapportages te gebruiken. Voor de groep docenten en leerlingen van de scholen die mee hebben gedaan aan dit deelonderzoek kunnen we concluderen dat volgens hen meerdere factoren een rol spelen in het feit dat de DTT rapportages vaak nog niet zijn gebruikt. Toch zijn er ook docenten die beschreven dat zij de DTT rapportages wel hebben gebruikt om vervolgstappen te nemen. Dit beschrijven we bij de derde onderzoeksvraag.

Uit de gesprekken die de UM gevoerd heeft met twaalf Limburgse schoollocaties blijkt tevens dat het ‘hoe nu verder’ onvoldoende duidelijk was uit de DTT rapportage. De DTT faciliteerde de docenten nog onvoldoende in het toepassen van differentiatie in de klas. Het instrument zou een concrete uitkomst moeten geven, bij voorkeur leerlingen automatisch moeten indelen in groepen en bijpassende vervolgstappen en oefenmateriaal moeten bieden. Daarnaast blijkt uit de gesprekken dat betrokkenen in de school onvoldoende het nut en de toegevoegde waarde van de DTT zagen. Er werd al van andere instrumenten gebruik gemaakt, de DTT kwam daar boven op en gaf niet altijd extra informatie. Bovendien werd de DTT niet omarmd door een deel van de mensen in de school die erbij betrokken waren. Het besluit tot deelname aan de pilot was vaak een besluit vanuit de directie of het bestuur. Vertrouwen van docenten (en leerlingen) in het instrument is echter minstens zo belangrijk om het tot een succes te maken.

3) Welke *best practices* zijn er op de verschillende niveaus rondom de DTT te identificeren?
Uit het UT-deelonderzoek naar de DTT rapportages blijkt dat drie van de twaalf docenten vervolgstappen hebben genomen op basis van de DTT rapportages. Dit zijn docenten van verschillende scholen waarbij er twee lesgeven in 3 havo en één lesgeeft in 2 vmbo tl. Eén docent gaf aan dat ze leerlingen eerst uitlegde wat de DTT was en dat ze daarna met iedere individuele leerling besprak welke diagnoses ze hadden gekregen op de verschillende onderdelen van de DTT. Daarna gaf ze de leerlingen voorbeelden van mogelijke oorzaken van onder-niveau diagnoses en vertelde ze leerlingen dat ze meer aandacht moeten besteden aan de onderdelen waarop ze een onder-niveau diagnose hadden. Daarnaast zocht deze docent naar websites die leerlingen konden gebruiken om te oefenen met spelling en herhaalde ze instructie over kommagebruik en aanhalingstekens. Ook beschreef ze dat ze leerlingen mondelinge feedback gaf over spelling wanneer leerlingen met opdrachten bezig waren tijdens de les. Een andere docent gaf aan dat hij de groepsrapportage met de leerlingen had besproken, door de resultaten met een beamer op een scherm te projecteren, te laten zien dat op het onderdeel ‘getallen en variabelen’ de meeste onder-niveau diagnoses waren, en te vragen wat zij dachten dat hier de oorzaak van was. Daarna liet hij leerlingen oefenen met opdrachten over getallen en variabelen, zoals het uitrekenen van de wortel van een getal en het oplossen van vergelijkingen. De derde docent gaf ook aan dat hij de groepsrapportage met de leerlingen had besproken, door de resultaten te projecteren en de leerlingen te laten reflecteren op de resultaten. Vervolgens benadrukte deze docent bij de leerlingen dat veel leerlingen een onder-

niveau diagnose hadden op het onderdeel woordenschat en dat het belangrijk is dat hun woordenschat wordt vergroot. Aangezien slechts enkele docenten en geen leerlingen vervolgstappen hebben genomen op basis van de DTT rapportages richten we ons nu op de vierde en tevens laatste overkoepelende onderzoeksvraag.

De meeste scholen uit het UM-onderzoek geven aan dat het prettig is om de uitslag op de DTT, als zijnde een methode-onafhankelijk en landelijk genormeerd instrument, te vergelijken met bevindingen uit andere instrumenten. De DTT vormt zo een extra gegeven naast andere informatiebronnen. Bovendien kan aan de hand van de DTT onderzocht worden of de lat binnen andere instrumenten in de school wel op de juiste hoogte ligt. Daarnaast gaf een school aan dat zij werken met wekelijkse maatwerkuren. Tijdens de maatwerkuren zijn docenten aan de slag gegaan met enkele leerlingen die veel aandachtspunten hadden uit de DTT rapportage. Een andere school gaf aan op directieniveau te werken met een dashboard waarin de schooldoelen in het kader van kwaliteitszorg gemonitord worden. Uit instrumenten die de school gebruikt, zoals de DTT tijdens de pilot, wordt voornamelijk die informatie gehaald die bijdraagt aan het dashboard. Door de data op deze manier te kanaliseren voorkomt de school het risico van een teveel aan data, waardoor er juist niets mee gedaan wordt.

4) Hoe kan gestimuleerd worden dat de DTT voor meer scholen, klassen, afdelingen en groepen leerlingen relevant en bruikbaar is?

Op basis van de factoren die docenten en leerlingen hebben genoemd in het UT-deelonderzoek naar de DTT rapportages kunnen we een aantal aspecten noemen om te stimuleren dat de DTT, en data in het algemeen, gebruikt worden door docenten en leerlingen om hun onderwijs en leren te verbeteren (zie tevens Schildkamp & Poortman, 2015). We zoomen in op drie aspecten. Een eerste aspect is dat de attitude tegenover het gebruiken van data om het onderwijs te verbeteren verder kan worden ontwikkeld. De docenten en leerlingen uit ons deelonderzoek gaven bijvoorbeeld aan dat de rapportages niet altijd door de docent en leerling zijn ontvangen of dat de schoolleider het gebruiken van de DTT rapportages niet had verplicht. Een positieve attitude van schoolleiders, docenten en leerlingen tegenover het belang om toetsen, die reeds zijn afgenomen bij leerlingen, door docenten en leerlingen te (laten) interpreteren en op basis daarvan vervolgstappen te nemen is belangrijk. Een tweede aspect is dat de kenmerken van de rapportage verder kunnen worden ontwikkeld. De docenten en leerlingen uit ons deelonderzoek gaven bijvoorbeeld aan dat de DTT rapportages niet op item-niveau laten zien of en welke fouten er zijn gemaakt door de leerlingen, dat de rapportages geen adviezen bevatten over welke vervolgstap docenten en leerlingen kunnen nemen en dat terminologie in de rapportages niet altijd duidelijk is. Een derde aspect is dat de kennis en vaardigheden van docenten en leerlingen verder kunnen worden doorontwikkeld om data, zoals data uit de DTT rapportages, te kunnen gebruiken om hun onderwijs en leren te verbeteren. Professionalisering rondom het formatief gebruik maken van data is een manier om dit zowel docenten als leerlingen aan te leren (Kippers, Poortman, Schildkamp, & Visscher, 2018).

In de gesprekken met scholen uit het UM-onderzoek zijn een aantal factoren benoemd die de bruikbaarheid en relevantie van de DTT kunnen vergroten. De belangrijkste factoren worden hier op een rij gezet. Allereerst het punt dat in de overkoepelende onderzoeksvraag 2 al aan de orde is

gekomen: zorg ervoor dat de DTT een concrete uitkomst heeft en er goede vervolgstappen geboden worden, waardoor de docent en de leerling weet hoe verder te gaan met de uitslag van de DTT. Bovendien is het van belang dat de rapportages eenvoudig te lezen zijn. Ten tweede zou het mooi zijn als de DTT onderdeel uitmaakt van een compleet meetsysteem voor de kernvakken in de onderbouw. Dit systeem geeft informatie op leerling-, groeps- en schoolniveau. Als onderdeel van het meetsysteem zou een instrument als de DTT gekoppeld moeten kunnen worden aan andere instrumenten (zoals methodes, LVS) zodat passende vervolgacties meteen voor handen liggen. Ten derde blijkt er behoefte te zijn aan een instrument wat op verschillende momenten afgenomen kan worden (flexibel inzetbaar). Daarnaast zou het in delen opnieuw afneembaar moeten zijn, zodat leerlingen sommige onderdelen opnieuw kunnen testen. De DTT sluit zo beter aan bij individuele leerroutes van leerlingen. Tot slot is het van belang dat het instrument omarmd wordt door alle lagen in een school. Betrek scholen bij de ontwikkeling van het instrument aan de voorkant, en zorg dat het instrument op alle onderdelen (bv organisatie, toets, rapportages) professionaliteit en vertrouwen uitstraalt.

Tabel 2. Docentkenmerken UT-onderzoek

		N	(%)
Geslacht	Man	4	33.3%
	Vrouw	8	66.7%
Leeftijd (tijdens dataverzameling)	≤ 30 jaar	5	41.7%
	31-40 jaar	2	16.7%
	41-50 jaar	4	33.3%
	≥ 51 jaar	1	8.3%
Opleiding (hoogst genoten)	Wetenschappelijk onderwijs	2	16.7%
	Hoger onderwijs	10[37]	83.3%
	Middelbaar onderwijs	0	0.0%
Functie	Docent in leerjaar 2	9	75.0%
	Docent in leerjaar 3	3	25.0%

Tabel 3. Leerlingkenmerken UT-onderzoek

		N	(%)
Geslacht	Jongen	4	33.3%
	Meisje	8	66.7%
Leeftijd	13	4	33.3%
	14	3	25.0%
	15	3	25.0%
	16	2	16.7%

[37] Drie docenten beschreven tijdens het interview dat ze nog in opleiding waren tot bevoegd docent.

Leerjaar	2	9	75.0%
	3	3	25.0%
Opleidingsniveau	Vmbo	9	75.0%
	Havo	3	25.0%
	Vwo	0	0.0%
Diagnose op DTT[38]	Onder niveau	1	8.3%
	Onder en op niveau	2	16.7%
	Onder, op, en boven niveau	1	8.3%
	Op en boven niveau	7	58.3%
	Boven niveau	1	8.3%

Conclusies

Wat betreft de vraag voor welke scholen, afdelingen, klassen en groepen de DTT de *meeste relevantie en bruikbaarheid* heeft, zien we dat de meeste respondenten uit zowel het UM als het UT onderzoek de DTT nog niet of nauwelijks gebruikt hebben. Het gaat hierbij om een variatie aan docentkenmerken wat betreft geslacht, aantal jaren leservaring, vooropleiding en leerjaar en schoolkenmerken wat betreft grootte en regio. Uit de bevindingen op vraag 3 blijkt dat 3 docenten van verschillende scholen en met verschillende achtergrond de rapportages wel hebben gebruikt. Het gaat bij zowel UM als UT echter om kleinschalig onderzoek, waardoor het lastig is om conclusies over de bruikbaarheid naar school, groep, afdeling, klas en groep te trekken. Wel heeft het onderzoek meer inzicht geleverd in de *factoren* die een rol hebben gespeeld bij het gebruik. Docenten en leerlingen hebben de rapportage nog niet altijd onder ogen gekregen. De rapportages laten geen fouten op item-niveau zien en geven geen advies over vervolgstappen. Ook aspecten zoals de perceptie van nut en toegevoegde waarde of de verplichting om het te gebruiken speelden een rol. Enkele docenten hebben de rapportages wel gebruikt. In deze *best practices* zien we dat deze docenten de rapportage klassikaal besproken hebben en verder hebben toegelicht, ook wat betreft het actie ondernemen voor verbetering van het leren. Één van de docenten heeft de resultaten ook individueel met de leerlingen doorgenomen. De DTT-bevindingen worden ook als vergelijking gebruikt voor andere instrumenten: omdat het methode-onafhankelijk en landelijk genormeerd is kan het als aanvullend en als ijking gebruikt worden. Een school zet het ook op individueel niveau in voor maatwerkuren en een school gebruikt het voor een dashboard in het kader van de kwaliteitszorg. Om gebruik van de DTT te *stimuleren*, zijn de attitude tegenover datagebruik voor onderwijsverbetering, de kenmerken van de rapportage en de kennis en vaardigheden voor formatief datagebruik van belang. Bovendien zou het mooi zijn als de DTT onderdeel uitmaakt van een compleet meetsysteem voor de kernvakken in de onderbouw, met informatie op leerling-, groeps- en schoolniveau. Een koppeling met andere instrumenten (zoals LVS) om beter inzicht te krijgen in vervolgacties is tevens belangrijk. Ook flexibel inzetten en erkenning van het belang van het instrument zouden de bruikbaarheid bevorderen.

[38] Diagnose op aspecten van Engels schrijfvaardigheid, Nederlands schrijfvaardigheid of wiskunde: (1) alle aspecten onder niveau, (2) sommige aspecten onder en sommige aspecten op niveau, (3) sommige aspecten onder, sommige aspecten op en sommige aspecten boven niveau, (4) sommige aspecten op en sommige aspecten boven niveau, of (5) alle aspecten boven niveau.

Referenties

Beckers, J., & Verstegen, D. (2016). Haal meer uit groepjes! *Didactief, 46*(9), 24-25.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*(1), 5.

Bosker, R. (2005). *De grenzen van gedifferentieerd onderwijs*. Groningen Universiteit Groningen.

Cito (2012). Diagnostische tussentijdse toets Verslag van de voorstudie. Cito Arnhem.

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review, 24*(2), 205-249.

College voor Toetsen en Examens (2014), Publieksversie toetswijzer diagnostische tussentijdse toets voor Nederlands, Engels en wiskunde. Technisch rapport. CvTE Utrecht.

Commissie Parlementair Onderzoek Onderwijsvernieuwingen (2008). Tijd voor Onderwijs. Eindrapport. Kamerstukken II, 2007-2008, 31 007, nr. 6.

Coubergs, C. (2014). *Binnenklasdifferentiatie: leerkansen voor alle leerlingen*: Acco.

Denessen, E. (2017). Verantwoord omgaan met verschillen: sociale-culturele achtergronden en differentiatie in het onderwijs.

Deunk, M. I., Doolaard, S., Smalle-Jacobse, A., & Bosker, R. J. (2015). *Differentiation within and across classrooms: A systematic review of studies into the cognitive effects of differentiation practices*: GION onderwijs/onderzoek, Rijksuniversiteit Groningen.

Fluckiger, J., Vigil, Y. T. y., Pasco, R., & Danielson, K. (2010). Formative feedback: Involving students as partners in assessment to enhance learning. *College teaching, 58*(4), 136-140.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81-112.

Heemskerk, I., Verbeek, F., Kuiper, E., Oomens, M., & Hilbink, E. (2014). Opbrengstgericht werken in het voortgezet onderwijs. In: Amsterdam/Utrecht: Kohnstamm Instituut/Oberon.

Hoogland, I., Schildkamp, K., van der Kleij, F., Heitink, M., Kippers, W., Veldkamp, B., & Dijkstra, A. M. (2016). Prerequisites for data-based decision making in the classroom: Research evidence and practical illustrations. *Teaching and teacher education, 60*, 377-386.

Keuning, T., van Geel, M., Frèrejean, J., van Merriënboer, J., Dolmans, D., & Visscher, A. J. (2017). Differentiëren bij rekenen: een cognitieve taakanalyse van het denken en handelen van basisschoolleerkrachten. *Pedagogische Studiën, 94*, 160 - 181.

Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention?. Studies in Educational Evaluation, 56, 21-31.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education, 31*(2), 199-218.

OCW (2011), Actieplan Beter Presteren: opbrengstgericht en ambitieus Het beste uit leerlingen halen

Onderwijsinspectie. (2008). *De staat van het onderwijs*. Utrecht: inspectie van het onderwijs.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and teacher education, 26*(3), 482-496.

Schildkamp, K., & Poortman, C. (2015). Factors influencing the functioning of data teams. Teachers college record, 117(4), 1-42.

Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. (2013). Toetsen met leerwaarde. Een reviewstudie naar de effectieve kenmerken van formatief toetsen. *Formative assessment. A review study on characteristics of formative assessment, NWO, Den Haag*.

Tomlinson, C. A., Brimijoin, K., & Narvaez, L. (2008). *The differentiated school: Making revolutionary changes in teaching and learning*: ASCD.

Van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. (2015). Integrating data-based decision making, Assessment for Learning and diagnostic testing in formative assessment. *Assessment in Education: principles, policy & practice, 22*(3), 324-343.

van der Wel, J., Paulussen-Hoogeboom, M., & Dekker, B. (2017). *Monitor pilot DTT: Bevindingen eindmeting*. Amsterdam: Regioplan.

Visscher, A. J., & Ehren, M. (2011). *De eenvoud en complexiteit van Opbrengstgericht Werken*: Universiteit Twente, Vakgroep Onderwijsorganisatie en-management.